DOCUMENT RESUME

ED 321 556 FL 018 675

AUTHOR	Tallmadge, G. Kasten; And Others
TITLE	Bilingual Education Evaluation System. User's Guide, Volume I, Recommended Procedures and Volume II, Technical Appendices.
INSTITUTION	RMC Research Corp., Mountain View, Calif.
SPONS AGENCY	Department of Education, Washington, DC.
PUB DATE	Nov 87
CONTRACT	300-85-0140
NOTE	228p.; For related documents, see ED 291 650, FL 018 677, and FL 018 631.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	Achievement Tests; Administrator Guides; *Bilingual Education Programs; Classroom Observation Techniques; Data Collection; Data Processing; Elementary Secondary Education; *Evaluation Methods; Interrater Reliability; *Outcomes of Education; Program Effectiveness; *Program Evaluation; Regression (Statistics); Research Design; Standardized Tests; *Statistical Analysis; Technical Writing; *Testing Problems; Test Reliability; Translation
IDENTIFIERS	*Bilingual Education Evaluation System; Extrapolation

ABSTRACT

The Bilingual Education Evaluation System was developed to help local bilingual education projects overcome evaluation obstacles and design sound, useful evaluations within the constraints of available funding and with responsiveness to current federal legislation and regulations. The evaluation system proposed involves a process component, an outcome component, and procedures for integrating them. An innovative element is a gap-reduction design for assessing student outcomes. This element assesses the academic growth of project participants relative to one of two recommended comparison groups: national norms or non-project students in the same grade. The user's guide to the evaluation system consists of two volumes. The first volume contains all of the system's procedures and practices. The system is divided into nine steps, each corresponding to a chapter of this volume: assuring that the project is evaluable; planning the evaluation; documenting program processes; selecting/adapting/developing instruments for assessing student outcomes; collecting outcome data; implementing the gap-reduction design; processing and analyzing data; integrating and interpreting results; and preparing evaluation reports. The second volume contains explanations and discussions of the rationale underlying recommendations for procedures and practices made in the first volume, and detailed guidelines on how to perform certain tasks. Topics include the following: classroom observation; interrater reliability; functional level testing; translating tests into other languages; major publishers of nationally standardized achievement tests; test reliability; quasi-experimental designs; gap reduction calculations; extrapolation procedures; and correcting for regression. (MSE)

ED321556

Bilingual Education Evaluation System

Users' Guide

Volume I - Recommended Procedures

November 1987

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

---

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

FL018675

BEST COPY AVAILABLE

RMC Research Corporation -- 2570 El Camino Real West, Mountain View, CA 94040

BILINGUAL EDUCATION EVALUATION SYSTEM

USERS' GUIDE

Volume I

Recommended Procedures

G. Kasten Tallmadge
Tony C. M. Lam
Nona N. Gamel

November 1987

Prepared for:

U.S. Department of Education
Washington, D.C.

By

RMC Research Corporation
2570 West El Camino Real
Mountain View, CA 94040

1-50-88-10

The research reported herein was performed pursuant to Contract No. 300-85-0140 with the U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official U.S. Department of Education position or policy.

Table of Contents

Page

List of Tables.... v
List of Figures.... v
Acknowledgments.... vi
I. INTRODUCTION.... 1
  Design Objectives for the BEES.... 2
  Nature of the BEES.... 3
  Overview of the *Users' Guide*.... 4
II. ASSURING THE PROJECT'S EVALUABILITY.... 7
  Is the Needs Assessment Adequate?.... 8
  Are Project Goals and Objectives Adquately Formulated?.... 11
  Is the Project Design Adequately Described and Consistent with Objectives?.... 14
  Are the Evaluation Questions Adequately Defined?.... 17
III. PLANNING THE EVALUATION.... 19
  Form an Evaluation Team.... 20
  Additional Planning Activities.... 24
IV. COLLECTING PROCESS DATA.... 29
  Specifying the Data to be Collected.... 30
  Selecting Data Collection Methods.... 37
  Developing Instruments for Collecting and Recording Data.... 40
  Preparing for Data Collection.... 45
  Collecting and Recording Process Data.... 46
V. SELECTING/ADAPTING/DEVELOPING INSTRUMENTS FOR ASSESSING OUTCOME OBJECTIVES.... 49
  Content Validity 50
  Assessing Content Validity.... 61
  Reliability.... 64
  Cultural Bias.... 66
  Summary of Test-Selection Recommendations.... 67

VI. COLLECTING OUTCOME DATA 69
- Planning the Testing Schedule 69
- Training the Test Administrators 71
- Preparing the Students for Testing 72
- Preparing the Testing Setting 73
- Administering Achievement Tests 74
- Collecting Other Test Data 75
- Collecting Non-Test Data 75
- Training Collectors of Non-Test Data 77

VII. IMPLEMENTING THE GAP-REDUCTION DESIGN 79
- Selecting a Comparison Group 80
- Assessing Gap Reduction Using Non-Test Data 84
- Assessing Gap Reduction Using Test Scores 86
- Using Medians Rather than Means 91
- Dealing with Regression Biases 93
- Assuring Representativeness of the Data 97
- Going Beyond the Gap-Reduction Design 99

VIII. PROCESSING AND ANALYZING DATA 101
- Analyzing Student Outcomes 101
- Analyzing Project Processes 113

IX. INTEGRATING AND INTERPRETING RESULTS 119
- Interpreting Gap Reductions and RGIs 119
- Drawing Conclusions and Making Recommendations 128

X. REPORTING 131
- Executive Summary 132
- The Main Body of the Report 133
- Checklist of Mandated Reporting Requirements 137

LIST OF TABLES

Page

Table 1 Suggested Tasks for Members of the Evaluation Team........ 25
Table 2 Examples of Self-Report Items... 42
Table 3 Minimum Requirements for Non-Test Data........................... 76

* * * * *

LIST OF FIGURES

Page

Figure 1 Interaction of project design and evaluability assurance........ 9
Figure 2 Sample chart for scheduling an evaluation................................ 23
Figure 3 Illustration of gap reduction.. 81
Figure 4 Graphic display of outcome and process data........................... 127

ACKNOWLEDGMENTS

The authors of this *Users' Guide* are significantly indebted to a large number of individuals who reviewed one or more of its multiple drafts and provided valuable comments and suggestions. We wish to give special thanks to the following reviewers of an early draft.

George R. Burket--California Test Bureau
Thomas D. Cook--Northwestern University
Joy A. Frechtling--Montgomery County (MD) Public Schools
Lupe A. Gonzales--Mission (TX) Independent School District
Raymond Morales--La Joya (TX) Independent School District
Elizabeth R. Reisner--Policy Studies Associates
Gerald Richardson--Florida State Department of Education
Kim O. Yap--Northwest Regional Educational Laboratory

Reviewers of a later draft, to whom we are also most grateful, included:

Gary J. Echternacht--Educational Testing Service
Joy A. Frechtling--Montgomery County (MD) Public Schools
Susan L. Reichman--RMC Research Corporation
Elizabeth R. Reisner--Policy Studies Associates
Jane L. David--Bay Area Research Group

We spent a year field testing the *Users' Guide* at nine school districts around the country. The site persons who gave their time most generously during this effort are listed below. Others too numerous to name, contributed both to the field test and to revisions of this document. We thank them all sincerely.

- *Tucson (AZ) Unified School District*
  - *Kathy Escamilla*
  - *Olivia Arrieta*

- *San Jose (CA) Unified School District*
  - *Margaret Payne Graves*
  - *Kim-Anh Nguyen*

- *Valley Center (CA) Union Elementary School District*
  - *Sarah Clayton*

- *Polk County (FL) School Board*
  - *Patricia A. Murrell*
  - *Mary H. Topping*

- *New York City Community School District Five*
  - *Victoria Manero*
  - *Gilda M. van Sand*

- *New York City Community School District Ten*
  - *Margery R. Falk*

- *Salem-Keizer (OR) Public Schools*
  - *Cheryl K. Crawley*
  - *Sally Edmiston*
  - *Susan Haverson*

- *Tyler (TX) Independent School District*
  - *Rena McGaughey*
  - *Henrietta Grooms*

- *Valley View (TX) Independent School District*
  - *Fernando Castillo*

We also thank:

J. Michael O'Malley--Eastern Region Evaluation Assistance Center
Robert J. Bransford--Western Region Evaluation Assistance Center

for their insightful comments and suggestions on recent drafts of the *Users' Guide.*

We are especially indebted to James J. English of ED's Office of Planning, Budget and Evaluation who served as the government's Project Officer throughout the entire research and development effort. His diligence in pursuit of excellence contributed immeasurably to the quality of this, the final product.

Finally, we owe thanks to the following Department of Education staff who also read various drafts of this document and provided valuable input:

John Chapman--Office of Planning, Budget and Evaluation
Edward Fuentes--Office of Bilingual Education and Minority Language Affairs
Alan Ginsburg--Office of Planning, Budget and Evaluation
Patricia Johansen--Office of Bilingual Education and Minority Language Affairs
Valena W. Plisko--Office of Planning, Budget and Evaluation
Nancy Rhett--Office of Planning, Budget and Evaluation

Not all of the comments we received were consistent with one another, and even the authors themselves are not in total accord on all points. We made changes in response to every comment we received but those changes often differed somewhat from the reviewers' exact suggestions. Compromises were required to reflect different viewpoints. The senior author was the final arbitrator of all such compromises and is thus responsible for whatever failures to achieve a balanced presentation remain.

GKT

I. INTRODUCTION

Title VII evaluations have been required since 1968.

Since the Bilingual Education Act (Title VII of the Elementary and Secondary Education Act) was passed in 1968, thousands of projects have received Federal funds to help them address the needs of limited-English-proficient students--students whose native languages and cultural backgrounds separate them from mainstream students.

Projects funded under the original Act were required to conduct self-evaluations, and some form of evaluation requirement has been in force--almost without interruption--from 1968 to the present time.

Most early evaluations were seriously flawed.

As a result of the legislation and associated regulations, a very large number of evaluation reports have been produced with Title VII funds. Unfortunately, most of these evaluations have not proven very useful either in helping local projects to demonstrate their effectiveness or in helping them to improve their services. Some of the reasons for these problems have been:

- lack of evaluation expertise at the local level.
- inadequate guidelines for evaluation.
- insufficient technical assistance to local projects.
- limited availability of funds.

What led to development of the BEES?

The Bilingual Education Evaluation System (BEES) described in this *Users' Guide* was developed to help local projects overcome these problems and design sound, useful evaluations within the constraints of available funding. Its development was sponsored by the Department of Education, which also established two regional Evaluation Assistance Centers (EACs) to help local projects improve their evaluations.

Design Objectives for the BEES

What objectives were established for the BEES?

Development of the BEES was guided by three primary design objectives:

- The system should reflect the sum total of knowledge gained from previous work in bilingual education evaluation. To meet this objective, the *Guide's* authors conducted a literature review which is summarized in a separate document.[1]

- The system should be useful at the local level for purposes of project improvement. To meet this objective, evaluation has been made an integral part of the project development process--from needs assessment, through project planning and implementation, to project modification.

---

[1] Tallmadge, G. K., Lam, T. C. M., & Gamel, N. N. (1987). *The evaluation of bilingual education programs for language-minority limited-English-proficient students: A status report with recommendations for future development*. Mountain View, CA: RMC Research Corporation.

- The system should be totally responsive to the current federal legislation and regulations governing the evaluation of Title VII projects.

Nature of the BEES

What is encompassed by the BEES?

The BEES is a total evaluation system that involves a process component, an outcome component, and procedures for integrating the two. The most innovative element of the system, however, is the *gap-reduction design* that is recommended for assessing student outcomes.

The gap-reduction design is recommended for assessing outcomes.

In developing the BEES we assumed that most projects would find it difficult or impossible to implement a traditional true or quasi-experimental design. For this reason, we formulated a design (the gap-reduction design) that is easy to implement, satisfies the regulations' requirements, and does not require a nonproject comparison group made up of students *similar* to those served by the project.

The gap-reduction design assesses the academic growth of project participants relative to one of two recommended comparison groups--national norms or the nonproject grade mates of the project students. These comparisons provide useful information about how well the students are progressing--but without yielding a quantitative estimate of the project's impact. The evaluator must rely on a wide array of information and findings about both processes and outcomes to form judgments about project effectiveness.

The BEES emphasizes quality control and explicit acknowledgment of unavoidable problems as methods of avoiding erroneous conclusions.

Overview of the *Users' Guide*

Content and organization of the Users' Guide.

The *Users' Guide* is organized into two volumes. This first volume contains all the system's evaluation procedures and practices. Volume II contains explanations and discussions of the rationales underlying various recommendations, as well as detailed guidelines on how to perform certain tasks described in Volume I.

The Bilingual Education Evaluation System includes nine major steps, each of which corresponds to a chapter of Volume I of the *Users' Guide*.

- Assuring that the Project is Evaluable (Chapter II).
- Planning the Evaluation (Chapter III).
- Documenting Program Processes (Chapter IV).
- Selecting/Adapting/Developing Instruments for Assessing Student Outcomes (Chapter V).
- Collecting Outcome Data (Chapter VI).

- Implementing the Gap-Reduction Design (Chapter VII).
- Processing and Analyzing Data (Chapter VIII).
- Integrating and Interpreting Results (Chapter IX).
- Preparing Evaluation Reports (Chapter X).

II. ASSURING THE PROJECT'S EVALUABILITY

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where--" said Alice.
"Then it doesn't matter which way you go," said the Cat.

(Alice in Wonderland)

What is evaluability?

The preceding quotation is directly relevant to the evaluability of projects. To paraphrase, if you don't know what a project is trying to accomplish--and why and how--you can't determine how successful it is, or how to improve it.

Evaluability implies the potential for meaningful and accurate evaluation which can lead to valid conclusions about project effectiveness and to sound recommendations for project improvement. A project is evaluable if its rationale, design, and expected outcomes are clearly defined and logically consistent. Assuring the project's evaluability is the first step in the evaluation process.

Evaluability assurance should be a collaborative endeavor.

Assuring evaluability and planning for the evaluation should be collaborative efforts involving the project director, the evaluator, and (if possible) the teaching staff. This group should review needs assessment data, descriptions of project goals and objectives, and planned project activities while addressing the following questions:

Four evaluability issues.

- Is the needs assessment adequate?
- Are the project goals and objectives adequately formulated and appropriate to needs?
- Is the project design adequately described and consistent with goals and objectives?
- Are the evaluation questions adequately defined?

A flow diagram representing the process of assuring the evaluability of a project is displayed in Figure 1. In the remainder of this chapter, we provide some guidelines for addressing the four issues to be considered in assuring evaluability.

Is the Needs Assessment Adequate?

What constitutes an adequate needs assessment?

Needs assessment is the process of:

- identifying the academic achievement and other school-related deficits of the target students.
- establishing a baseline against which to compare students' performance after they have received project services.

The current (1986) regulations specify that achievement must be assessed in English language proficiency. That term is used broadly to denote proficiency in speaking, reading, writing, and/or understanding English. Proficiency must also be

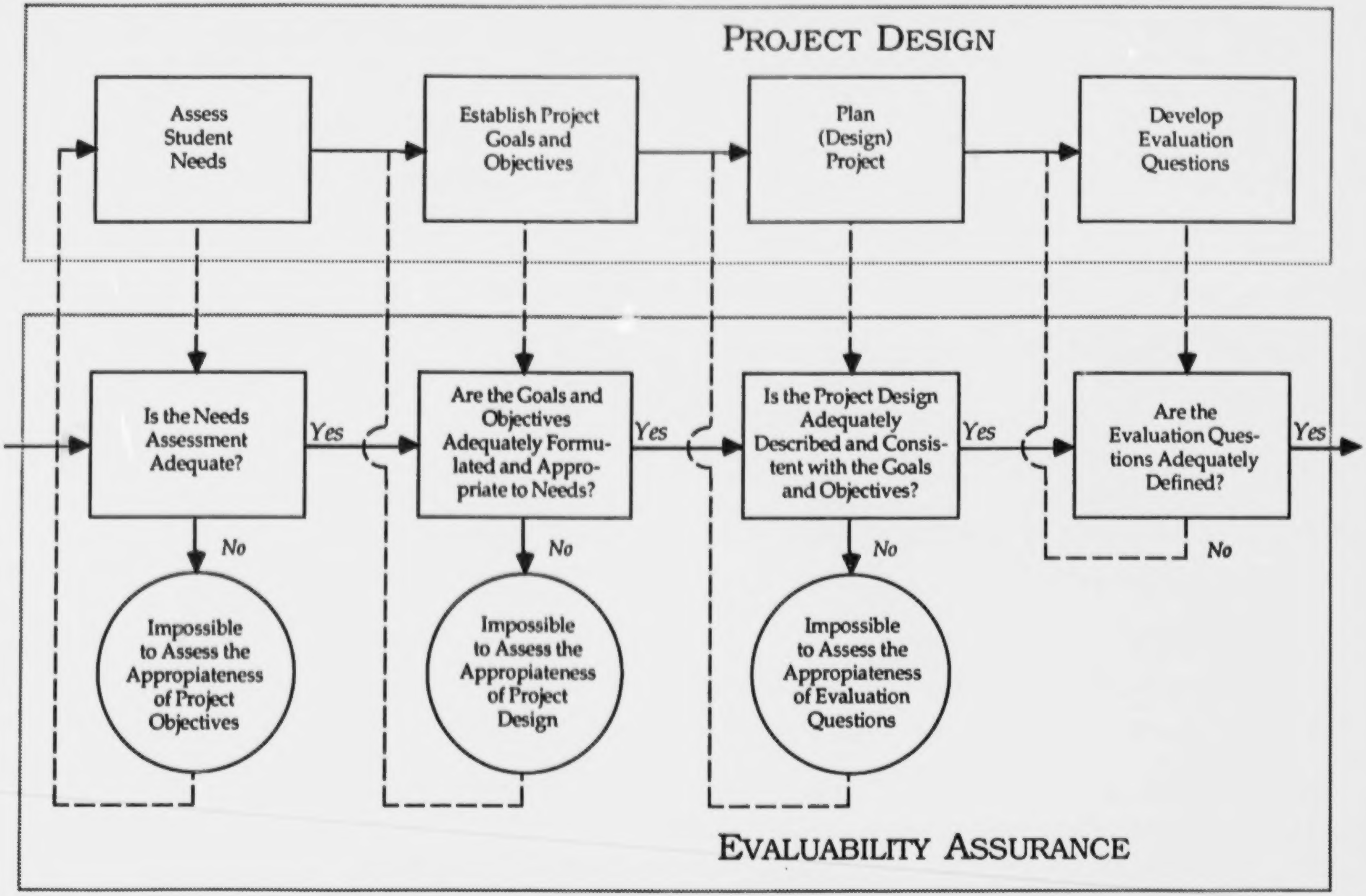

PROJECT DESIGN
Assess Student Needs
Establish Project Goals and Objectives
Plan (Design) Project
Develop Evaluation Questions
Is the Needs Assessment Adequate?
Yes
No
Are the Goals and Objectives Adequately Formulated and Appropriate to Needs?
Yes
No
Is the Project Design Adequately Described and Consistent with the Goals and Objectives?
Yes
No
Are the Evaluation Questions Adequately Defined?
Yes
No
Impossible to Assess the Appropiateness of Project Objectives
Impossible to Assess the Appropiateness of Project Design
Impossible to Assess the Appropiateness of Evaluation Questions
EVALUABILITY ASSURANCE

Figure 1. Interaction of project design and evaluability assurance.

assessed in the students' native language (for projects of developmental bilingual education) and other courses or subjects of study. Reliable and objective assessment methods must be used in all of these assessments according to the current bilingual education regulations [§501.31(a)(1&2)].

English language proficiency measures.

For the assessment of needs in English language proficiency, it would be desirable to have as many of the following indicators as possible and/or appropriate:

- home language survey.
- English language proficiency test.
- standardized English reading readiness or reading test.
- teacher evaluations.

Native language proficiency measures.

For the assessment of native-language proficiency, it would be desirable to have as many of the following indicators as possible and/or appropriate:

- home language survey.
- native-language proficiency test (if available).
- teacher/aide evaluations.
- records of prior schooling.

Academic subject matter measures.

For the assessment of subject matter knowledge in other academic content areas, it would be desirable to have as many of the following indicators as possible:

- standardized achievement test in written English.
- standardized achievement test written in L1.
- other achievement tests in L1.
- records of prior schooling.

Different students may have different needs.

If the target students constitute a homogeneous group, needs may be appropriately characterized using means, medians, or other descriptive statistics. On the other hand, where the group to be served is heterogeneous in terms, for example, of ethnic origin, literacy in L1, or amount of prior schooling, it will be appropriate to make a careful note of the different needs of different subgroups.

If the already collected needs assessment data are not complete or adequate, you must work out a strategy for collecting and documenting adequate information about student needs.

Are Project Goals and Objectives Adequately Formulated?

Project objectives must relate to student needs.

Outcome *goals* for projects may be no more than restatements of identified student needs. If, for example, the needs assessment found that the target LEP students were two years behind their mainstream peers in some academic area, the corresponding project goal would be to reduce or eliminate that achievement gap.

Goals and objectives should determine project design.

The primary purpose of adequately formulated goals and objectives is to facilitate design of the intervention that will address the identified needs. A goal as broad as reducing the English language proficiency gap is not, however, particularly useful for that purpose. Broadly stated goals need to be broken down into their component *objectives*. To illustrate this point, the goal of gap reduction *might* be broken down into objectives such as:

- developing literacy skills in L1.
- transferring L1 literacy skills to English.
- enhancing the English vocabularies of target students.

Each of these objectives could profitably be broken down into still more specific objectives. Literacy skills in L1 (or English) might include, for example, recognizing the letters of the alphabet, knowing letter-sound correspondences, being able to decode letter combinations and words, etc. The greater the level of detail, the more useful the objectives will be in the project design process. The one essential ingredient of a well formulated objective is that it relates clearly to an identified student need. Going back to our example, while some educators might debate the effectiveness of teaching decoding skills, those skills are clearly relevant.

Objectives must be measurable.

All objectives must be stated in measurable terms so that you can subsequently determine whether they have been achieved. "Enhancing students' appreciation of music" is an example of an inadequately stated objective since appreciation of music can mean many different things to different people. If, on the other hand, you were to say, "music appreciation as measured by such and such a test," you would have solved the problem by making the objective measurable. Some people might not agree that the test really measured music appreciation, but it would at least be clear what you hoped to impart to the students.

Stating objectives in measurable terms does not imply that you must establish a quantitative level of performance as a criterion of success. Keeping up with, or catching up to, mainstream students will serve quite well as a criterion for success. There is no need to specify that the gap between the two groups' performance levels should be reduced by a specific number of units. In fact, establishing that kind of criterion is inappropriate given the state of our current knowledge about bilingual education. We simply do not know what would be reasonable to expect.

The linking of objectives to needs.

When reviewing project objectives, you should always verify that each one is logically linked to an identified student need. If it is not, then the objective should be discarded.

Statements of objectives for bilingual projects should also reflect whatever diversity of needs exists within the target group. With heterogeneous groups, different objectives must be developed for each subgroup that has unique needs.

Multi-year projects require annual objectives.

It is also necessary for evaluability assurance purposes that separate objectives be developed for each year of a project. Otherwise, the required annual evaluations might show failure to achieve an objective that was not relevant to that year's instructions. If, for example, instruction in English reading is not introduced until the second project year (after literacy in L1 has been achieved), it would be inappropriate to include gains in English reading achievement as a first-year objective.

If the project objectives do not meet the criteria discussed above, they must be modified or reformulated. Without sound and well written goals and objectives, you will not be able to assess the appropriateness of the project plan.

Is the Project Design Adequately Described and Consistent with Objectives?

All project features should be described in detail.

There must be a detailed and complete description of the project before a useful evaluation can be undertaken. That description should encompass every topic to be covered in the curriculum, the classroom hours that should be devoted to it, and the instructional practices that should be followed in teaching it. In addition, anticipated activities related to teacher training, curriculum development, and parent involvement must be described.

Design features should be linked to project objectives.

The first thing to consider when evaluating a project's design is whether its linkages to the stated outcome objectives are adequately supported by logic or empirical evidence. In essence, you need to determine that the curriculum, teaching strategies, and other project characteristics "make pedagogic sense" given the project's objectives. To do so, you should draw upon the research literature (on effective instructional practices, second language acquisition, linguistics, etc.) and your own experience and expertise.

If you find that some curriculum component, teaching strategy, or other project characteristic is unrelated to one or more of the stated objectives, that project feature should be deleted and perhaps replaced with a more relevant one.

Not only must project features be clearly relevant to outcome objectives, they must be described in sufficient detail so that classroom observation can subsequently determine the extent to which they are present in the project *as implemented*.

Key features must be identified.

The project design should spell out what should be going on in the classroom at any given time in terms of both content and method of presentation. This information is essential so that the observer knows what he or she should look for during the observation period and will be able to recognize discrepancies between what was intended and what is actually happening. Features should also be prioritized so that the observer's attention can be appropriately focused.

Design features become process objectives.

Features discussed in the project design--particularly key features--are traditionally called *process objectives*. The task of determining the extent to which they are present in the project as implemented is traditionally called *process evaluation*.

Each item in a detailed project description can be regarded as one process objective. For example, the statement that kindergarten students will receive instruction in L1 on the concepts of up-down, left-right, long-short, near-far, and inside-outside, for three hours a week using the XYZ science series is a description of how the project is intended to operate. Hence it is a process objective.

If the project plan is not adequately described, you must work to revise it so that there is no confusion about how the project should operate. Chapter IV presents some guidelines for project description and documentation.

Are the Evaluation Questions Adequately Defined?

Evaluation questions should derive from project objectives.

Evaluation questions should be a natural outgrowth of your project's outcome and process objectives. In fact, each objective can be turned into an evaluation question simply by asking the extent to which it was achieved. If the outcome objective was to reduce the gap between project and mainstream students in reading achievement (as measured by a particular test), an appropriate outcome-evaluation question would be, "How much was the gap reduced?" If the process objective was for teachers to conduct all instruction in English, an appropriate process-evaluation question would be, "What percentage of the instruction was conducted in English?"

Additional questions will derive from the regulations.

In addition to the evaluation questions that derive directly from process and outcome objectives, you will have to consider questions posed by the current evaluation regulations *whether or not* they relate to project objectives. For example, your project might not have objectives related to student absenteeism. Nevertheless, because you are required to report data on absenteeism, your evaluation must address the question, "Has the project affected the rate of student absenteeism?"

Prioritize your evaluation questions.

Because of resource and budget limitations, you may not be able to answer every evaluation question you would like to ask about your project. For this reason, it is important that you prioritize your questions giving precedence to those that derive directly from the regulations and those that relate to key project features.

In addition to process- and outcome-related questions, you should *always* address questions that pertain to data validity. Even if you can't afford to collect or analyze data related to the validity questions, you should think about them throughout the evaluation period and acknowledge possible problems when interpreting your evaluation findings.

Once the entire process depicted in Figure 1 has been successfully completed, the project's evaluability will have been assured, and planning for specific evaluation activities can begin. Evaluation planning is the topic of the next chapter.

III. PLANNING THE EVALUATION

Plan carefully to avoid problems.

Evaluation planning is the process of developing an orderly and efficient blueprint for collecting, processing, analyzing, interpreting, and reporting the data needed to address the evaluation questions that have already been developed. An evaluation that is not carefully planned is almost certain to encounter problems that proper planning can avoid. Among other things, proper planning can ensure that:

- all the needed data are collected.
- no unnecessary data are collected.
- the data are as free from errors as they can be.
- project staff and school personnel will cooperate with the evaluation.
- results are useful to project staff and available early enough to be helpful.

First, study the Users' Guide.

Before you begin planning your evaluation, you should thoroughly familiarize yourself with the contents of the *Users' Guide* and make sure that the project's evaluability has been verified as described in Chapter II.

Then, start planning as early as possible.

Planning should begin as soon as possible after grant award. An even earlier start would be highly desirable since the time available to complete a few of the tasks (e.g., selecting, ordering, and obtaining standardized tests) is barely adequate even with the six-month preservice period that is now an integral part of Title VII basic grants.

Ideally, all evaluation planning activities should be completed during the six-month preservice period. If that is not possible, you must at least complete those planning tasks associated with data collection since actual data collection should begin at the same time as service delivery.

Form an Evaluation Team

Include persons with diverse skills.

To assure the best possible planning, you should set up an evaluation team that includes at least the project director, the project evaluator, and one or (preferably) more project teachers. While the project director is the appropriate leader, the size and makeup of the evaluation team may vary as a function of the size of your project, district resources, and other considerations. It is important, however, that all of the following competencies be represented:

- methodological expertise in evaluation, measurement, data management, and statistical analysis.
- knowledge of bilingual education theory and practice.
- knowledge of the project's curriculum and instructional strategy.
- interpersonal and communication skills.

Form action committees.

The project director should organize the evaluation team into committees with specific responsibilities for planning and carrying out all of the required tasks. Initially, there should be committees for (a) selecting tests, (b) developing strategies and instrumentation for collecting process evaluation data, and (c) developing procedures and instrumentation for documenting the project's student, treatment, and setting characteristics.

Involve teachers in selecting tests.

The *test-selection committee* should be made up of teachers and/or curriculum developers who have detailed knowledge of the project's specific instructional objectives. If a person with psychometric training is available, he or she would be a valuable member of the committee--especially during the test-selection process.

Have project designers plan the process evaluation.

The *process-evaluation committee* should include the project director, the project evaluator, and anyone else who was involved in designing the project and has first-hand knowledge of how the project was intended to be implemented.

Include a person who knows the record-keeping systems.

The *documentation committee* should include one or more persons familiar with all record-keeping practices and procedures from the classroom to the district level.

Members of each committee should study those parts of the *Users' Guide* that are relevant to their responsibilities before any evaluation planning is begun. Then, as new planning or implementation tasks are undertaken, new committees should be formed, and additional review of *Users' Guide* recommendations should be conducted.

Develop a calendar of evaluation activities.

The first meeting of the evaluation-planning team should be devoted to developing a calendar of events associated with planning for and conducting the evaluation. Figure 2 presents a sample format for such a calendar. You should use it as a starting point for developing your own. Whatever format you ultimately decide upon, your evaluation plan should encompass all of the following activities:

- selecting, modifying, and/or developing outcome evaluation instruments.
- developing instruments for collecting data on student, project, and setting characteristics.
- training data collectors and collecting data.
- scoring and coding data.
- setting up a data management system.
- creating and editing data files.
- analyzing data.
- interpreting analytic findings.
- preparing the evaluation report.

All of these topics are covered in considerable detail in the following chapters of this *Users' Guide*.

Activity	Resp. Staff	Start Date	End Date	Completed?	Remarks
Outcome Measures					
Review Candidate Instruments					
Select and Order Tests					
Train Test Administrators					
Administer Pretest					
Administer Posttest					
Records Extraction					
Develop Instrument(s)					
Train Data Extractors					
Extract Data					
Teacher Reports					
Develop Instrument(s)					
Train Interviewers					
Conduct Interviews					
Classroom Observation					
Develop Instrument(s)					
Train Observers					
Conduct Observations					
Data Processing					
Develop Coding Procedures					
Train Coders					
Code Data					
Exercise Quality Control					
Data Analysis					
Develop Data Analysis Plan					
Set Up Data Management System					
Create Computer Files					
Edit Files					
Conduct Analysis					
Report Preparation					
Prepare Draft Report					
Obtain Reviewer Comments					
Revise Report					
Submit Report to Interested Audiences					

Figure 2. Sample chart for scheduling an evaluation.

Assign specific tasks to team members.

In addition to working out a detailed schedule for evaluation activities, assigning responsibilities for completing those activities is an important part of the planning process. You should note that a column for recording personnel assignments is included in the sample schedule shown in Figure 2.

Many, if not most, evaluation activities will be performed by the project director, the project evaluator, and project teachers. Still others may be performed by aides, parents, and district administrative and clerical staff. If possible, all of these individuals should participate in at least those planning sessions where their possible roles are discussed.

Involvement breeds cooperation.

Involving as many people as possible in planning the evaluation--particularly teachers and administrators--will be well worth whatever extra effort it entails. Involvement leads to a sense of ownership, and ownership leads to tolerance and cooperation when evaluation activities become intrusive (as they invariably do). Table 1 lists some of the tasks in which various members of the evaluation team should participate.

Additional Planning Activities

Good planning should not stop with the mechanics of setting up schedules, forming committees, and making personnel assignments.

Table 1

Suggested Tasks for Members of the Evaluation Team

Project Directors

- assuring the project's evaluability
- reviewing the evaluation plan
- assigning tasks to members of the evaluation team
- coordinating the evaluation team
- selecting, modifying, or developing instruments
- arranging and monitoring data collection and processing
- ensuring that evaluation tasks are completed on time
- conducting classroom observations
- assisting in interpreting results
- reviewing, editing, and writing parts of the report

Evaluators

- assuring the project's evaluability
- developing or refining the evaluation design
- selecting, modifying, or developing instruments
- conducting classroom observations
- managing and analyzing the data
- interpreting the data
- writing the evaluation report
- presenting evaluation findings

Teachers and Paraprofessionals

- assisting in assuring the project's evaluability
- reviewing the evaluation plan
- selecting, modifying, or developing instruments
- collecting and processing data
- assisting in interpreting the data

Parents and Community Members

- reviewing the evaluation plan
- suggesting evaluation questions

School District Administrators

- reviewing the evaluation plan
- suggesting evaluation questions
- visiting classes informally

District Evaluation and Testing Staff

- reviewing the evaluation plan
- providing district data
- assisting in the construction of the project's data base
- reviewing draft reports

Good planning also includes: (a) smoothing the way for activities that could otherwise encounter resistance and (b) anticipating potential problems (and at least thinking about contingency plans).

Overcoming resistance to evaluation.

Smoothing the way has already been discussed briefly in the context of involving teachers and administrators in evaluation planning and implementation. Involvement and communication are the two most important keys to smooth operations. You should:

- solicit evaluation questions from all persons who have a stake in the project (including parents).
- disseminate information about the project and how evaluation can lead to its improvement.
- inform all concerned parties well in advance of scheduled evaluation activities that may require their time or disrupt their schedules.
- provide teachers and administrators with prompt feedback regarding test scores.
- train all data collectors to be as nonintrusive as possible.
- perform quality-control checks on all data collectors.
- be responsive to all questions and criticisms.
- report all findings in a timely manner.

Anticipating potential problems.

The remaining chapters of this *Users' Guide* include descriptions of many of the kinds of problems you may encounter in implementing your evaluation.

You should read these discussions carefully as problems often *do* arise unexpectedly, and you need to recognize them quickly and respond appropriately. Inappropriate or delayed responses could threaten the validity of an entire year's evaluation.

This *Users' Guide* has been developed in the belief that the evaluation of bilingual education projects can accomplish much more than a minimal satisfaction of legislative requirements. Thoughtful planning and careful attention to evaluation practices are necessary, however, if evaluation results are to be useful in improving the education of language minority students. The remaining chapters of the *Guide* will help the evaluation team plan appropriate evaluation activities.

IV. COLLECTING PROCESS DATA

Purposes of process data.

Section 500.51 of the current evaluation regulations requires that you collect data on student, project, and staff characteristics. We recommend that you also collect data on any events that may affect project outcomes. These data will serve two purposes:

- documenting discrepancies between what was intended (process objectives) and what actually occurred.
- documenting project and context characteristics for scientific and accountability purposes.

The necessity of collecting data for scientific and accountability purposes is easily understood. Information related to process objectives, however, should be directly useful to you in working toward project improvement.

Full implementation is unlikely in first years.

Research has demonstrated that a new project is rarely fully implemented in its first years. Some teachers may be following the project's curriculum quite closely, but others are likely to be deviating substantially from intended practices. Still other teachers might be following project plans as far as using L1 to teach math, but might be inadequately proficient to use L1 exclusively in teaching science.

Deviations from the original project design can be adaptive or maladaptive. In either case, outcome data will be easier to interpret if you know, for example, that science was taught partially in English rather than exclusively in Vietnamese or that L1 language arts textbooks did not arrive until January. Outcome data cannot be used for program improvement unless actual project circumstances and events are known.

Steps in collecting process data.

There are four steps to the collection of process data:

- specifying the data to be collected and reported.
- selecting sources and methods for collecting the data.
- preparing to collect the data.
- collecting and recording the data.

Each of these steps is explained in more detail below.

Specifying the Data to be Collected

You will be collecting data to meet the requirements of the regulations and for internal use in your project. While there will be some overlap between these two types of information, we will discuss them separately.

Required data on student characteristics.

Student characteristics data. The current regulations require that you collect data on *the educational background, needs, and competencies of the limited English proficient persons served by the project.* [§ 500.51(a)]

The regulations do not specify exactly what data must be collected, but we consider the following list to be minimal:

- age.
- grade level.
- first language.
- ethnicity.
- language used in the home.
- proficiency in L1 (including literacy).
- proficiency in English.
- socioeconomic status (e.g., participation in National School Lunch Program).

Ideally, these data should be collected for each student and reported using summary statistics.

If necessary, describe community characteristics.

If certain data (e.g., socioeconomic status) are too difficult to collect for individual students, and you are serving a homogeneous group, a description of the community might suffice. For example, "students served by our program come from a community where over half the families receive Aid to Families with Dependent Children. Most of the students' parents have had fewer than five years of schooling and are not literate."

Additional data that may be useful.

There are other data that would be useful to you in planning your project or modifying it to increase its effectiveness. Such information is also useful to audiences of your report because it helps to explain why your project is designed as it is. These data might include:

- prior education, including years of schooling and possibly type (e.g., monolingual English classroom).
- parents' literacy in L1.
- parents' proficiency in English.

You should strive to collect all student characteristic information that will be useful to readers of your report in understanding your project. If project students have characteristics that are central to the project design, but are not listed here (such as giftedness) these characteristics should certainly be included even though they exceed minimal reporting requirements.

Project characteristics data. The current regulations require that you collect data on *the amount of time (in years or school months...) participants received instructional services in the project and, as appropriate, in another instructional setting.* [§ 500.50(b)(3)(ii)(A)]

Required data on length of services.

The minimal data needed to satisfy this requirement are:

- date of entry into the project.
- date of exit from the project.
- (if project participation is for less than full time or less than a school year) other instructional settings in which the student is served, such as mainstream classroom, special education, Chapter 1, etc.

These data should be collected for each student and reported using summary statistics.

Required data on educational activities.

The regulations also require that you collect data on:

- *The specific educational activities undertaken pursuant to the project.* [§ 500.51(b)]
- *The pedagogical materials, methods, and techniques utilized in the program.* [§ 500.51(c)]
- *With respect to classroom activities, the relative amount of instructional time spent with students on specific tasks.* [§ 500.51(d)]

The minimal data needed to satisfy these requirements are:

- subject areas included in project instruction.
- major curriculum objectives.
- a listing of project materials different from those used in mainstream classrooms.
- percentage of instructional time devoted to L1 language arts.
- instructional content areas taught in L1.

- length of time students are expected to remain in the project.
- for each major curriculum objective, the hours per year at each grade level devoted to the objective.

To illustrate the final item above, you might report for an objective such as *Accurate addition of single-digit numbers*, "0 hours in kindergarten, 36 hours in first grade, 40 hours in second grade, 20 hours in third grade."

Additional data on educational activities.

Additional data may be necessary to describe your project accurately, such as the use of aides, peer tutoring, parent education, or computer-assisted instruction.

Staff characteristics data. The current regulations require that you collect data on *the educational and professional qualifications, including language competencies, of the staff responsible for planning and operating the project.* [§ 500.51(e)]

Required data on staff characteristics.

We consider the following items to be the minimum that will satisfy the requirements:

- level of education.
- credentials and certificates.
- bilingual/bicultural teaching experience.
- other teaching experience.

- languages understood and spoken, including degree of fluency, and whether English is spoken with a heavy accent.

Data on project implementation. In addition to the process data you must collect to satisfy the regulations, you will also want to collect data that will tell you whether your project is operating as originally intended. If you find that certain activities are not taking place as planned, you will have to determine whether the changes represent an improvement or whether the project would work better as it was originally designed.

Collect data to check on process objectives.

Your project's process objectives should tell you what, if any, additional process data should be collected for use in your evaluation. Ideally, you would collect data to verify that each objective had been met. Realistically, you may have to review your objectives for critical project features or activities that are not covered in the minimal reporting requirements listed in the previous section. Thus, if a process objective specified that students will be grouped in classes according to their L1 literacy, you would want to collect information on student grouping. Then an examination of students' literacy in L1 (data that must be collected anyway to satisfy the regulations) will tell you if students were grouped as planned.

Areas of possible data collection.

Each project will have its own process objectives and critical features. It is not possible to specify here what these features will be. However, some areas to keep in mind while thinking about critical activities or objectives are:

- parent involvement.
- staff development.
- materials development.
- patterns of classroom language use.
- use of aides, resource teachers, or other staff.
- use of special equipment or materials.
- integration of students' home cultures into classroom activities.
- use of specific classroom management techniques.
- methods of project management or coordination.

Questions to determine important project features.

As an additional help in determining what project features you will want to document, you can address these questions:

- In what ways do you expect a project classroom to differ from a mainstream classroom at each grade level?
- What special or additional services does your project provide that are not normally provided by the school district?

- What aspects of the project are intended to meet needs that are unique to the target population?

Plan ahead for data interpretation.

In order to interpret student outcome data, you will need to know the degree to which the project has been implemented as intended in critical areas. Therefore, you must plan to document critical processes even if the documentation goes beyond that required by the regulations.

Selecting Data Collection Methods

Consider three data collection methods.

Three general data collection methods are discussed in this section. They are:

- examination of records.
- self-reports by project staff.
- observation of project activities.

You will probably want to use some combination of these methods to collect process data.

If possible, use existing records.

Examining records. Once you have determined what data you want to collect, you should investigate the possibility that the data have already been recorded somewhere in district, school, or classroom files. Characteristics of students and qualifications of staff are two reporting areas that probably have extensive documentation you can use. Some project activities will also be documented in classroom records, although in most cases you will find that record-keeping is not uniform across teachers.

Potentially useful records.

Records which can provide useful process information are:

- student files.
- teachers' daily activity logs.
- lesson plans.
- records of homework assignments.
- records of special assistance provided to certain students.
- minutes or agendas from parent advisory council meetings.
- in-service training announcements or schedules.
- project personnel resumes and/or applications.

It may also be possible for teachers to create or augment certain records in ways that will be useful for evaluating project processes. They could, for instance, be asked to include language of instruction.

Self-reports can be questionnaires, interviews, or written reports.

Self-report measures. You can ask project staff for information about activities not recorded in any existing document. Self-reports can be reliable sources of data, but they are likely to be biased if the questions are value-laden (e.g., Did you do what you were supposed to do?). Self-reports, as long as they are not self-evaluations, can be very useful sources of information--particularly if they can be supported by other data (such as classroom observations).

There are three methods of collecting self-reports.

- Staff can write periodic reports of what they did (e.g., a monthly report of the percent of time lessons were taught in L1).
- Questionnaires can be given to staff to fill out (e.g., a questionnaire every semester about courses or workshops attended).
- Staff can be interviewed in person (e.g., a year-end interview investigating how information about students' home cultures was incorporated into classroom instruction).

There is no substitute for classroom observation.

Classroom observation. Formal or informal classroom observation should be done to verify self-reports and to document classroom processes. In order to determine whether the project is being implemented as planned, there is no substitute for a visit to the classroom. At a *minimum*, you should informally visit classrooms during times when specific subjects are being taught.

Plan for even informal observation.

Before each visit, you should ask yourself what you expect to see. Should students be using locally-developed materials? Should the teacher be using

English with a carefully selected vocabulary? Should a team teacher be previewing instruction in L1? Should an aide be translating instruction for students who are having difficulty?

If you see exactly what you expected to see, there is reason to believe the project is being implemented as intended. If not, you should find out why. Possibly the project as designed was not working for the students, the teacher, or both. Possibly, the teacher needs more training or more time to become comfortable with the project. To assist in determining why discrepancies are observed, some form of self-report may be useful.

Developing Instruments for Collecting and Recording Data

For even the most minimal data collection effort, you will need to develop forms to record and store your data. You will need a form for each student and for each instructional staff member in order to record data required by the regulations. The forms should be of your own design, depending on the data you decide to collect. Make sure, however, that forms are clearly labeled and dated and that each student record includes the student's full name and district identification number.

Design a form for each record or purpose.

Forms for record extraction. If you will be extracting process information from records such as agendas of parent advisory council meetings or teachers' lesson plans, you will need to design a form for each type of record. Make sure that each form is clearly labeled and dated, and if someone else will be filling in the form, that it includes complete instructions.

Forms for self-report data. To develop an instrument to collect self-report data, you will need to decide:

- the type of self-report you want (written report, questionnaire, or interview).
- whether you want to collect data periodically or only once per year.
- whether you will have time and money to analyze individual written or oral responses or whether you want to construct closed-choice questionnaires.

Brief questionnaires can be easy and economical.

Table 2 contrasts the types of questions suitable for each type of report. For economy of data collection, processing, and analysis, a brief, closed-choice questionnaire is the best choice. This type of questionnaire requires careful construction, however, to make sure that the response alternatives you offer reflect everything your respondents would like to say. The first time you use such a questionnaire, it is best to allow respondents to write in answers if they wish. This will enable you to see if your choices need to be revised.

Table 2

Examples of Self-Report Items

Type of Instrument		
Self-Report	Questionnaire	Interview
1. Describe how you used your classroom aide this term during English language arts instruction.	How did you use your classroom aide this term during English language arts instruction? Rank the three most frequent activities (1 = most frequent). ___correct paperwork ___translate for class ___work with small groups ___clerical and administrative work ___instruct class ___handle discipline problems	How did you use your classroom aide this term during English language arts instruction? Probes: Did you use the aide to help with ________? Why or why not?
2. Compare your use of L1 and English this year for instruction in science.	At the beginning of the year, what language(s) did you use for instruction in science? ___mostly L1 ___half L1 and half English ___mostly English At the end of the year, what languages did you use for instruction in science? ___mostly L1 ___half L1 and half English ___mostly English	How did your use of English and L1 for instruction in science change over the year? Why?

Types of observation instruments.

Forms for classroom observation. For even informal classroom observations, you will need a simple form on which to record your observations. If observations will be more formal or will be done by more than one person, you will need a more detailed instrument. There are three common types of observation instruments:

- behavioral checklists (on which the observer tallies the number of times he or she sees a specific behavior).
- coded behavior records (on which *sequences* of specific behaviors can be coded).
- delayed report instruments (which observers use to describe what they have seen over a short period of time).

Each of these observation methods is discussed briefly in Appendix A, along with its advantages and disadvantages. You should give each careful consideration before deciding which is likely to be most appropriate for your project.

Focus on key project features.

Regardless of the particular method of classroom observation that is employed, it is simply not possible for a single observer to see everything that goes on in a classroom. For this reason, the best strategy is to focus observations on those actors and activities that are most critical to the intended implementation of the project and to its success.

If the project designer has specified that all instruction should be conducted in English and has developed outcome expectations on the presumption that it will be, clearly the language of instruction must be a focus of classroom observations. Other examples might include strict adherence to a hierarchy of instructional objectives, or immediate positive reinforcement of all correct student responses. It should be noted, however, that features considered critical to the proper implementation of one instructional design (and therefore to its success) may be of little concern to another.

If a project's design has been well and fully explicated, it should be possible to identify the most critical things to observe. Even when this is the case, however, it would be good practice to review classroom observation plans with the project designer. With a less well explicated project design, such review is absolutely indispensable.

Avoid high-inference observations.

If, after some period of observation, you ask the observer to assess a classroom's "climate," you are going to get a highly subjective response and one that would not be consistent across different observers. One observer might describe it as "noisy, chaotic, and out-of-control", while another might see it as "individualized, spontaneous, and creative." To avoid this kind of subjective inconsistency you should make the observational task as objective as possible.

You will get better results by having your observer count the number of positive reinforcements made by the teacher in a 20-minute period than by having him or her rate the teacher on a 7-point scale from positively to negatively reinforcing. Even more reliable data will be obtained if the observer does no more than tally each time the teacher uses the word "good," since simply deciding what constitutes an instance of positive reinforcement requires a moderately high level of inference.

Multiple observers must agree on what they see.

If all classroom observations are done by a single observer, high inference observations are less problematic. When multiple observers are used, however, it is essential that they be trained to the point that they have a high level of agreement (interrater reliability) regarding what they see. Adequate interrater reliability (see Appendix B) will be *much* easier to obtain with low-inference observation tasks. No amount of training may produce adequate interrater reliability if the observation tasks require high levels of inference.

Preparing for Data Collection

Make sure data collectors know what to do.

Different data collection methods will require slightly different preparatory steps. In general, you will always need to:

- select and train data collectors (unless you plan to collect all the data yourself),
- check data collectors' early efforts to make sure data collection is done correctly.

For classroom observation, these activities are more complicated than for record extraction. For training observers, you will need a videotape of a class or a live nonproject class which can be observed for practice. You will also need to verify interrater reliability if more than one observer is used. (See Appendix B.)

Collecting and Recording Process Data

Once you have developed forms to collect all the data you need and you have trained your data collectors, data collection should be straightforward in most cases. For some projects, most data collection will be accomplished toward the end of the school year. In other instances, data will be collected throughout the year, e.g., as new students enter the project. If data are collected as events occur, they should be checked and consolidated every few months.

In collecting classroom observation data, there are additional points to remember:

- observe each classroom several times during the year.

- look for different features in different observation sessions--don't try to accomplish too much in one session.
- avoid scheduling observations at times when classroom processes will be atypical, such as near a holiday.
- make sure observers remain unobtrusive and do not disturb classroom activities.

Other events can affect project outcomes.

Recording non-project data. During the course of the school year, events may occur which affect project outcomes but which are unrelated to project objectives. Such events might include a teachers' strike, lengthy school closings because of bad weather, or a locally upsetting event such as a factory closing. By keeping in touch with project staff, you can find out about external events that are likely to affect student performance. You should keep a record of all such events, noting the facts of the events, the dates, and possible impact on students. You can refer to this written record when you interpret data.

V. SELECTING/ADAPTING/DEVELOPING INSTRUMENTS FOR ASSESSING OUTCOME OBJECTIVES

Poor test choices make projects look ineffective.

It is essential to pay careful attention to the task of choosing appropriate tests for evaluating bilingual projects. As we point out below, it is possible to make several different kinds of poor choices. Unfortunately, most poor choices will have the effect of making interventions appear to be less effective than they really are. Perhaps even more unfortunate is the fact that good test choices may be precluded by district or state rules and regulations.

With respect to the latter point, local-level evaluations *must* conform to the district and state regulations. If the required practices are unsound, however, and especially if they make projects appear less effective than they really are, tactful efforts to bring about district- or state-level change may have long-range benefits for all concerned parties.

Content validity is the most important selection criterion.

High reliability and validity are always considered to be desirable test characteristics, and they are specifically required by Section 500.50(b)(2)(ii) of the current evaluation regulations. Appropriate difficulty levels and freedom from cultural bias are also considered to be meritorious features. For our purposes, however, a particular form of validity--

content validity--will be of primary concern. As we shall see, a test that has high content validity will tend to have the other desirable characteristics as well.

Content Validity

Content validity means measuring the right stuff.

The content validity of a test is the extent to which that test taps the skills and knowledge it is supposed to measure. Content validity is always assessed by means of subjective impressions of item relevance.

In general, it can be said that a *pretest* has content validity if it accurately reflects the performance level of students who are about to enter a new instructional sequence. To do this, the test must measure both what students have already been taught and what they will be taught in the new sequence. It must measure both of these areas because some students will not have learned everything they were taught previously and others will already know some of what will be taught in the new sequence.

A *posttest* can be thought of as the pretest for the next instructional sequence. Thus, to have content validity, a posttest must measure what was taught in the previous instructional sequence as well as what will be taught in the next instructional sequence.

Content validity and difficulty are closely related.

At this juncture it is appropriate to point out that there is a close relationship between content validity and difficulty. If a test has high content validity, it should be neither too difficult nor too easy.

Since *too easy* and *too difficult* tend to have subjective meanings, we offer the following operational definitions:

Definitions of too difficult and too easy.

- A test is too easy if any of the testees know the answers to all of the questions. When this is the case, one can assume that some students would know the answers to additional, even more difficult questions.

- A test is too difficult if any of the testees do not know the answers to at least some of the questions. When this is the case, one can assume that some students would not know the answers to additional, even easier questions.

It is important to note that the preceding definitions use the phrase "know the answers" rather than "answer the questions correctly." Items may be answered correctly by random guessing, or answered incorrectly because of careless errors. Thus, tests can be too difficult even if there are no zero scores or too easy even if there are no perfect scores.

Assess the match of test items to curriculum content.

Our earlier statement--if a test has high content validity, it should be neither too difficult nor too easy--should not be taken to imply that tests which are neither too easy nor too difficult have high content validity. Content validity can only be assessed by comparing the test questions with the curriculum. If there is a good match, then there is high content validity.

Assess difficulty to verify content validity.

Looking at difficulty levels is a useful check on content validity--and if tests are found to be too easy or too difficult, something needs to be done to correct that problem--but checking difficulty levels is no substitute for a direct assessment of content validity.

Project evaluation imposes four content-validity requirements.

Before discussing procedures for assessing content validity, it is appropriate to examine in some detail the specific content-validity requirements you may have to address in evaluating bilingual education projects. Because the regulations specify that *the progress of project participants [must be] measured against an appropriate nonproject comparison group [§ 500.50(b)(1)]* evaluation instruments need to have content validity for four different data points:

- the pretest performance level of the project group;

- the pretest performance level of the comparison group;
- the posttest performance level of the project group;
- the posttest performance level of the comparison group.

Because comparison-group students may be substantially higher achievers than project students, these content-validity requirements may present difficult instrumentation problems--at least for some academic subjects.

Range of test coverage is a potential problem.

Even at the elementary grades, it may be impossible to find a single test that covers the range of performance levels from that of the project group at pretest time to that of the comparison group at posttest time. At the later grades, the problem will be much more severe. Especially in the area of English language proficiency, tests that are appropriate for one group may be totally inappropriate for the other--both in terms of content and in terms of difficulty.

Functional level testing is a potential solution.

There is only one adequate solution to this problem--testing the two groups with different *levels* of the same *test battery*. This strategy is called **functional level testing,** and it is a psychometrically sound method for measuring all four performance levels (see above) on a common scale. Appendix C presents a brief discussion of functional level testing for those who are not already familiar with the procedure .

In-level testing may make projects look ineffective.

It *may not* be necessary to resort to functional level testing if you can find a single test that has content validity for all four performance levels and is neither too easy nor too difficult. If you use a single instrument and it *does not* meet those requirements, however, it is almost certain that your project will be made to appear less effective than it really is. This problem is most likely to occur when an in-level (as opposed to a functional level) test is used for pretesting project students.

When project students are tested with a test that is too difficult, their scores are artificially prevented from reflecting their true performance levels. Raw scores of zero convert to the lowest performance level that the test is capable of measuring, while the students' true performance level may be substantially lower. The only way to determine the students' true performance is to administer a lower level of the test.

Note that a pretest performance level that is too high (because the true performance level was lower than the test could measure) results in a growth estimate that is too small. Again, under these circumstances, your project will be made to look less effective than it really is.

English language proficiency is broadly defined.

English language proficiency. English language proficiency is a term that the regulations and the *Users' Guide* use in the broadest possible sense to encompass understanding and speaking English, reading readiness, reading, and possibly even writing. It is thus not necessary that you employ a test labeled "language proficiency" for your evaluation. In fact, readiness and reading tests will usually serve evaluation purposes more effectively.

Language proficiency tests tend to have as their primary objective the classification of students as either LEP (needing bilingual services) or non-LEP (able to function adequately in mainstream classrooms). They are generally not well designed for project evaluation purposes and we recommend that you use them only when other instruments are *clearly* inappropriate.

Native language proficiency. Native language proficiency need be assessed only in projects of developmental bilingual education, although such assessment may be useful for other types of projects as well.

Suitable tests may be hard to find.

The main difficulty with assessing native language proficiency is the dearth of suitable instruments--especially for languages other than Spanish. Several English language reading and reading readiness tests have, however, been professionally translated into Spanish, Vietnamese, and other languages. Your regional (East or West) Evaluation Assistance Center should be able to help you identify these instruments.

For less commonly encountered languages, you may experience substantial difficulties in locating suitable tests. Only locally developed instruments may exist and they may be difficult to find. Again, however, one of the Evaluation Assistance Centers should be able to help.

If no suitable off-the-shelf tests can be found, you will have to develop your own or translate an English language test. Unfortunately, neither of these options is particularly desirable.

Translating tests is difficult and hazardous.

Translating tests is a very difficult task and one that is likely to be successful only with instruments covering very basic skill levels. Picture-word associations may qualify, but paragraph meaning questions will almost certainly present serious problems to even the most expert translators.

Translating from English into a non-Indo-European language will be particularly hazardous. You would probably be better off developing a test from scratch if you are working with Asian or Native American languages. Even with an Indo-European language, your chances of producing a successful translation will depend on the extent to which the

English version of the test has the following characteristics:

- short questions.
- active voice.
- specific rather than general terms.
- no metaphors or colloquialisms.
- no vague words (probably, frequently, sometimes).
- no subjunctive mood.

If the test you wish to translate has these characteristics, Appendix D provides some guidance on how to proceed.

Seek professional help in developing new tests.

As far as developing new tests is concerned, we recommend that you seek the help (not just the advice) of a professional expert in psychometrics. You should not consider using any instrument unless it has been item-analyzed and revised on the basis of that item analysis. This recommendation pertains to tests developed elsewhere as well as to any you may develop locally.

Compare progress in L1 to progress in English.

Assessing progress in native language proficiency *relative to a nonproject comparison group* appears to mean comparing the project group's progress in L1 with the comparison group's progress in English. This type of comparison requires either a test written in L1 that has national norms or a test without norms that has content validity for both the project group's L1 curriculum and the comparison group's English curriculum. This issue is discussed more fully under the following heading.

Suitable tests may again be hard to find.

Other academic subject areas. Achievement tests for assessing progress in non-language subject areas should be written in the language of instruction unless the students are more proficient in English. Again, there may be problems in finding tests that have content validity for the particular curriculum being taught *and* are written in the language of instruction. Your regional Evaluation Assistance Center is probably your best source of help.

An important consideration in selecting subject matter tests is how you will use them in your evaluation. You will need to quantify the pre- and posttest performance levels of both the project group *and* a comparison group. If you can find a suitable test that has national norms, you may use the 50th percentile of these norms as your comparison group.

Using norms with L1 translations of standardized tests.

Special equatings of the L1 and English versions of the test may have been done by the publisher, and special provisions may have been made for accessing the norms. Special norms may even have been developed for the L1 version of the test. Even if nothing special has been done, however, the gap-reduction design (see Chapter VII) will work quite well using the raw scores obtained by project students on the L1 version of the test and the raw scores corresponding to the 50th percentile of the appropriate English language norms. Whatever biases may exist in the pre- and posttest performance-level indicators cancel out when *relative growth* is what is being assessed.

Using live comparison groups requires curriculum comparability.

If you cannot find a suitable test with national norms, you will need to use an unnormed test with the mainstream grade mates of project students as your comparison group. Two conditions must be met, however, whenever mainstream grade mates are used as the comparison group:

- there must be no significant differences between the curricula being taught to the two groups.

- there must be both English and L1 versions of a test that has content validity for the common curriculum.

- If the two curricula are significantly different, you have only two choices. You may use an English language normed test (untranslated) or translate such a test into L1.

Using English language tests for subjects taught in L1.

If the project students have some proficiency in English (even though they are taught in L1), it may be possible to test them using an instrument written in English. Under these circumstances, it will probably be appropriate to make some modifications to the test so that it more accurately reflects the student's subject matter knowledge.

Three potential problems.

When tests are used with groups other than those for which they were designed, three kinds of problems are likely to be encountered:

- students will fail to understand what they are supposed to do.
- students will not have time to respond to all the items because of their slow reading.
- students will fail to understand some questions because of unfamiliar English words that are totally unrelated to the knowledge being tapped.

Three partial solutions.

All three of these problems are at least partially fixable. You can:

- Modify instructions to be sure they are clear to your project students. You may translate them *into* L1 or paraphrase them *in* L1.

- Extend time limits so that most students will have time to attempt most of the items on the test. This is just what the test developer had in mind when the time limits were established for non-LEP, mainstream students.

- Simplify the language of the test items--but only to the extent that the words or phrases simplified are independent of the content being tested. If you want to test the students' math skills, don't make them take a reading test in order to get to the math items.

Modifying tests may enhance reliability and validity.

Making the kinds of changes just discussed will actually enhance the reliability and validity of the test for your project students. It will only do so for raw scores, however. If the test has norms, they will not be valid for students taking the modified version of the test (although you can still use those norms as your comparison group in the gap-reduction design).

Modifying tests will also invalidate any interlevel equating. You should not use modified *out-of-level* tests with either of the recommended comparison groups (norms or mainstream grade mates).

Assessing Content Validity

As mentioned earlier, the content validity of a test is always assessed in terms of the extent to which its items tap the skills and knowledge the test is supposed to measure. It is not possible to make such an assessment from information provided by the test publisher or test information service centers. It is absolutely essential for persons familiar with your curriculum to examine candidate instruments on an item-by-item basis.

Identifying candidate instruments.

Before you begin the content validity assessment process, you must, of course, identify a group of candidate instruments. We recommend that you begin by contacting the major publishers of standardized achievement tests (see Appendix E) and obtain "specimen sets" of instruments that each publisher's sales representative feels would be appropriate for your target group. You should also contact your regional Evaluation Assistance Center for additional nominations. Finally, if it has not already been identified as a candidate, you should consider any instrument that is administered to your students for statewide or districtwide testing purposes.

Recommended procedures for assessing content validity.

While this task may appear quite burdensome, it should not be so in practice. In fact, it should be possible for a person who is familiar with the curriculum to perform a content analysis in approximately the same amount of time that person would require to take the test. The procedure we recommend is for the test evaluator to read each question and to mark on a regular IBM-type answer sheet whether:

- the item covers material that was taught to project and comparison group students prior to the project year that is just beginning (you may use answer alternative "a" to indicate such items).
- the item covers material that will be taught during the project year that is just beginning or about to begin (use answer alternative "b" to indicate such items).
- the item covers material that will be taught in years subsequent to the project year that is just beginning (use answer alternative "c" to indicate such items).
- the item covers material that is not included in the past, present, or future curricula (use answer alternative "d" to indicate such items).

Ideal distributions of item types.

Ideally, at pretest time, you would like to find that about 30% of the items were type "a", about 40% type "b", and 30% type "c". Type "d" items serve no useful purpose in your evaluation. Consequently, you would like to find very few or none of them.

At posttest time, the ideal distribution would be approximately 70% of the items being classified as types "a" or "b" and the remaining 30% as type "c".

Problems will result from missing item types.

Clearly, these ideal distributions will not be exactly matched by any existing tests. On the other hand, tests must have at least some items in categories "a", "b", *and* "c" in order to be well suited for evaluation purposes.

- Tests that lack type "a" items will be too difficult at pretest time and will yield performance-level estimates that are systematically too high.
- Tests that lack type "b" items will be insensitive to project-related learning.
- Tests that lack type "c" items will be too easy at posttest time and will yield performance-level estimates that are systematically too low.

Reliability

Test reliability is an extremely important consideration *when assessing individual students*. When assessing the progress of groups of students it becomes much less critical. Still, with typical, project-size groups, more reliable tests should be preferred over less reliable tests, all other things being equal.

Three relevant reliability considerations.

There are several reliability facts that you should consider:

- The reliability of a test is proportional to its length--but the length of a test is defined by the number of items that students respond to, not the number of items printed on its pages. A functional level test will be more reliable (and, as we have already discussed, more valid) than a test made up of items students can only guess at.

- Standardized achievement tests put out by the major test publishers are all about as reliable as such tests can be. Locally developed tests will usually be significantly less reliable.

- Test reliabilities change with the characteristics of the group tested. They will be significantly lower for the kind of homogeneous groups typically served by Title VII projects than for the much more heterogeneous norm groups for which published reliability coefficients were published.

Reliability is less important than content validity.

Appendix F provides some guidelines you can use to determine the effect of reliability differences on the accuracy of growth and gap-reduction estimates.

As that appendix demonstrates, however, reliability differences must be quite large or reliability levels quite low before the impact on the accuracy of growth estimates amounts to much. For the most part, reliability differences will be trivial compared to differences in content validity. Perhaps even more significant is the fact that tests with high content validity will tend to be more reliable than tests with low content validity. In the final analysis, you should probably not worry very much about test reliability.

Cultural Bias

Cultural bias has little effect on measures of growth.

Cultural bias, like reliability, is *extremely important* when assessing the status of individual students. When tests are used for evaluating bilingual projects, however, cultural bias is much less important. Whatever bias was present at pretest time will also be present at posttest time (although probably to a somewhat lesser degree). Thus, when pretest scores are subtracted from posttest scores, cultural bias tends to cancel out.

The most widely recognized form of cultural bias results from items that are more difficult for minority than for majority students. In extreme cases they may even have different "correct" answers. Most publishers of standardized achievement tests employ debiasing procedures to detect and eliminate such items.

While debiasing procedures are generally quite effective, they are all less than perfect. Debiased tests may still yield somewhat biased status indicators; the growth estimates they yield should be quite acceptable, however.

There is a second kind of cultural bias in tests--a bias that stems from mainstream students having acquired more test-wiseness than minority LEP students. This bias can be reduced by teaching test-taking skills to LEP students and instilling in them a more competitive attitude toward test-taking situations.

Gains in test wiseness confound growth estimates.

An increase in test wiseness will cause test scores to go up and cultural bias to go down. Note however, that an increase in test wiseness between pre- and posttesting will be confounded with whatever increase in subject matter proficiency occurred over the same period.

If project participants are observed to be catching up to their mainstream grade mates in terms of scores on an English language proficiency test, that catching up may be due partially or entirely to gains in test wiseness rather than gains in English language proficiency. You should keep this possibility in mind when interpreting your evaluation findings.

Summary of Test-Selection Recommendations

Use standardized achievement tests.

Whenever possible, we recommend that you use standardized achievement tests to evaluate bilingual education projects. While they have serious deficiencies when used for other purposes, they have no equal for measuring growth along well defined achievement dimensions. They also tend to be more reliable than locally developed tests.

Four additional recommendations.

Additional recommendations include:

- Use only tests that have high content validity for both the project group and the comparison group in your evaluation.
- Use functional level testing as appropriate to achieve high content validity.
- Whenever possible, avoid instruments labeled language proficiency tests. They were not designed for evaluation purposes and are not well suited for such usage.
- Use the mainstream grade mates of project students for your comparison group in preference to national norms whenever there is a good curriculum match and such usage does not impose an excessive testing burden.

VI. COLLECTING OUTCOME DATA

The regulations specify 12-month testing intervals.

The current evaluation regulations require that you collect both test and non-test data. The required test data are to be collected on a 12-month cycle and include: *objective measures of the academic achievement of [project] participants related to English language proficiency, native or second language proficiency (for programs of developmental bilingual education), and other subject matter areas* [§ 500.50(b)(2)(iv)].

Current and *former project participants must be tested.*

These data are to be collected from both LEP and native English speaking current project participants, and from former LEP project participants who were exited and are currently in mainstream classrooms.

To prepare for achievement testing you will need to plan the testing schedule, train the test administrators, and prepare the students and setting. Each of these steps is discussed below.

Planning the Testing Schedule

In scheduling tests, you should take the following important considerations into account:

Plan to avoid excessive testing.

- The regulations require a 12-month testing interval. If you use a standardized achievement test, you can avoid double testing by using each year's posttest as the following year's pretest, but you must select the best points at which to change test levels.

Spring-to-spring testing requires a separate baseline test.

- Usually, you will select either a spring-to-spring or a fall-to-fall schedule. If you select spring-to-spring testing, however, you will need to pretest all first year participants in the fall to establish a pre-project baseline performance level.

Norm-based comparisons impose special requirements.

- If you plan to use norms as your comparison group, the best testing time is within two weeks of the date the test was normed. With a 12-month testing interval, you *may*, however, use interpolated norms which the publisher will usually provide upon request.

Make provisions for testing (nearly) all project participants.

- The regulations require that evaluation findings be representative of the students served. This means that every effort must be made to test all students by:

(a) scheduling make-up testing sessions near the original testing dates to obtain scores for students who were absent.

b) scheduling special testing sessions for students who enter the project after the pretest or leave before the posttest so that their scores can be included in the evaluation. (If fewer than 10% of the project students are late enterers or early leavers, you may omit this procedure.)

Training the Test Administrators

All test administrators and proctors should be trained every year no matter how experienced they are. The training should emphasize the following points:

- Written instructions should be followed exactly, whether standardized or unstandardized tests are used. If there are no written instructions, you should prepare some to guarantee standardized test-administration practices.

Modified tests require written *administration instructions.*

- If modifications have been made to standardized test instructions, they must be written down and followed exactly at pretest *and* posttest times. No changes should be made between the two test administrations.

Watch for indications of invalid test scores.

- Administrators and proctors should watch for behaviors that may indicate invalid test scores. These behaviors include:

  (a) talking during the test.
  (b) looking at other students' answers.
  (c) marking answers at random.
  (d) finishing unusually early.
  (e) not paying attention to instructors.
  (f) leaving many answers blank.
  (g) leaving the room.
  (h) losing their place on the answer sheet.

Proctors or administrators should make a note if they observe any of these behaviors during testing. Student test scores judged to be invalid should be discarded.

Preparing the Students for Testing

Scores will be more valid if students are prepared for testing.

All students, but LEP students in particular, can benefit from thorough preparation for testing. Preparation should include:

- telling students when, where, and why they will be tested.

- helping students gain test-taking experience by allowing them to practice on similar tests and answer sheets, including any practice test that may be provided by the test publisher.

- teaching students test-taking skills such as guessing, saving difficult items until easier ones have been answered, budgeting time, etc.

- encouraging students to do their best on the test.

- notifying parents of the testing schedule so that they can help by making sure students are well rested and well nourished on testing days.

Preparing the Testing Setting

Try to find a clean, quiet, well lighted place for testing.

The testing setting can have a major impact on test scores. Ideally, the setting should have all of the following characteristics. Since our main concern is with change from pre- to posttest, however, the most important consideration is that the testing setting be *the same* at pre- and posttest times.

- The testing space should be clean, quiet, well lighted, and large enough to prevent crowding.

- The same space should be used for every test administration; pretesting, posttesting, and make-up testing.

- The space should be isolated from external noises such as softball games, band practice, police sirens, or lawn mowers.

- Noises such as a telephone, school bells, or the public address system should be silenced during testing.

- Test administrators and proctors should avoid noisy shoes or jewelry and should never whisper or converse during a testing session.

- Desks or tables should be large enough to hold both test booklets and answer sheets.

- Desks or tables should be far enough apart to discourage copying and to allow proctors to move around the room.

Administering Achievement Tests

Standardize testing conditions and procedures.

When tests are administered, it is important to maintain standard conditions for the pretest, posttest, and any make-up sessions in order to be sure your results can be interpreted. It is best to follow these suggestions:

- Always go over the practice examples with the students before beginning the test.

- Adhere strictly to administration directions and times.

- Note any unusual student behavior or any disruptions to the testing session.

Collecting Other Test Data

Make use of all available test data.

Although you are not required to by the regulations, you should collect any other test data that are available for your students. These include:

- scores from project entry and exit tests.
- scores from any state or district testing programs.
- scores from curriculum unit tests.
- scores from teacher-made tests.

Collecting Non-Test Data

Many unobtrusive measures are useful for evaluation.

Non-test indicators can be reliable and easy to obtain. Several such indicators are required (as appropriate) by the regulations. Table 3 describes these minimum data requirements.

We recommend that you also collect other, non-required data such as:

- student mobility rate.

Table 3
Minimum Requirements for Non-Test Data

Collect for Each Student	Report by Grade Level for Project Group
Years retained in grade as of (date).	Average retention rate range, # and % of target group with history of retention.
Whether or not student dropped out. (Not appropriate for elementary grades.)	#, % of target group (local definition of "dropout")
Number and percent of days absent from project classrooms broken into at least two or three calendar periods.	Average percent days absent, range.
Date of special education referral or placement.	#, % of target group referred or placed.
Date of placement in gifted and talented program.	#, % of target group placed.
Date of enrollment in post-secondary education institution. (Not appropriate for elementary grades.)	#, % of target group enrolled.

- number of disciplinary actions.
- number of suspensions.
- project teacher turnover rate.
- number of books checked out of school libraries by project students.
- numbers of times students were late to class.

These data can be used as behavioral indicators of affective states such as students' self concepts or attitudes toward school. We recommend that you use such behavioral measures instead of paper-and-pencil affective instruments. Such instruments are not sufficiently reliable or valid to use in assessing project effectiveness.

Before and *after data are needed.*

For non-test data, as for test data, you should attempt to obtain baseline data and then collect additional data at least annually. In some cases (such as for dropout rates) you may be able to obtain a state or national average for your students' ethnic group. This average can, if necessary, serve as your baseline figure.

Training Collectors of Non-Test Data

Data extractors also need to be trained.

Collectors of these data will abstract them from existing records (see Chapter IV for additional guides on examining records). They should be provided with record abstraction forms and taught:

- what data to collect.
- where to find the information.

- how to gain access to the information.
- how and where to record the data.
- what precautions to take to protect students' privacy.

Exercise quality control procedures.

Data collectors' work should be checked after they have extracted information on a few students. This check will show whether additional hands-on training is needed.

VII. IMPLEMENTING THE GAP-REDUCTION DESIGN

Gap reduction is a valid criterion for project success.

One primary goal of bilingual education is to close the gap in English language proficiency between project students and their non-LEP peers. A second primary goal is to keep project students from falling behind their non-LEP peers in other subjects while they are learning English.

It makes sense to talk about bilingual education projects and their success in terms of these gaps. Thus it is not entirely coincidental that we have developed an evaluation strategy called the gap-reduction design which we recommend for evaluating Title VII (and other) bilingual projects.

The gap-reduction design is recommended.

Our recommendation is not made lightly. We have critically examined all other designs described in the literature, the assumptions that underlie them, and the implementation requirements they impose. While three of those designs (see Appendix G) *may* yield better estimates of project impact if properly executed under ideal conditions, none was judged to be nearly as easy to implement or interpret as the gap-reduction design. This ease of implementation and interpretation stems from the design's focus on *achieving project objectives* rather than *quantifying the size of the treatment effect*.

The design is easy to implement.

Conceptually, the gap-reduction design is very simple. It involves only four, easy-to-measure quantities:

1. the project group's pretest performance level.
2. the project group's posttest performance level.
3. the comparison group's pretest performance level.
4. the comparison group's posttest performance level.

From these four quantities, we can easily measure the gap between groups at pretest time (#3 minus #1) and at posttest time (#4 minus #2). The amount of *Gap Reduction* then is simply the *Posttest Gap minus the Pretest Gap*. Figure 3 provides a graphic representation of these relationships.

Selecting a Comparison Group

Two comparison groups are recommended.

Before going on to discuss the procedures required to implement the gap-reduction design, it is appropriate to consider the issue of selecting a comparison group. There are two possibilities that we recommend:

- the mainstream grade mates of the project students.
- the 50th percentile of the national norms.

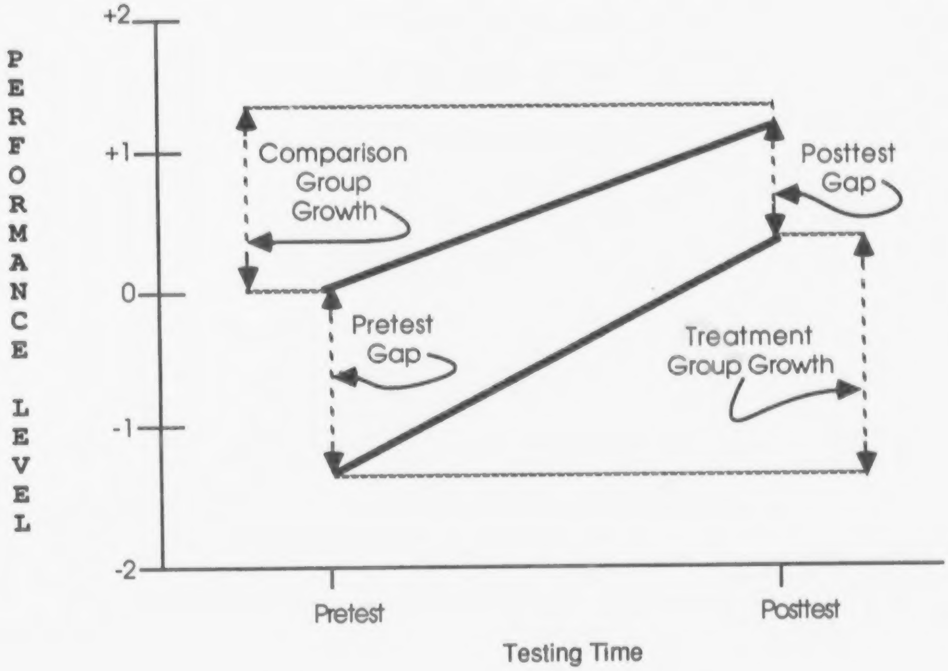


Figure 3. Illustration of gap reduction.

Using grade mates as a comparison group.

Using mainstream grade mates as your comparison group has several advantages.

- You have a local rather than a national basis for comparison--a factor that may facilitate interpretation of the results and make the comparison seem fairer.
- You are not restricted to standardized achievement tests (although you *may* use them).

- You have more freedom in selecting testing dates.

Using mainstream grade mates as your comparison group has only one potential disadvantage--but it may be a major one. You will have to collect data from all members of the comparison group. This requirement will not be too burdensome *if* you can "piggyback" your evaluation on an annual districtwide or statewide testing program. But if such piggybacking means you will be using a test with low content validity (see Chapter V), the end result will be to make your project appear less effective than it really is.

Using norms as a comparison group.

When we talk about using national norms, we are necessarily confining our attention to achievement testing. When the gap-reduction design is implemented with classroom grades, attendance figures, or a variety of other indices, it can only employ mainstream grade mates or some other "live" comparison group. Normative data that are statistically adequate to use as a substitute for a live comparison group are available solely for standardized achievement tests.

The main advantage of using norms for your comparison group is that you don't have to impose an additional testing burden on comparison-group students. The only disadvantages--and they may be of no concern whatsoever--are that you *must* use a standardized achievement test and you *should* test within two weeks of the test's empirical norming date.

Other comparison groups are possible.

Although we recommend using either the mainstream grade mates of your project students or the 50th percentile of the national norms as your comparison group, you *could* use a variety of other groups. One example would be a group of unserved LEP students from a neighboring school district. Many other possibilities exist.

If you use one or more the these "other" groups, your results will be difficult to compare with results from other projects. Interpretation of your results will also be more difficult. We urge all Title VII projects to use one of the two recommended comparison groups. Then, if you feel *additional* comparisons would provide useful information, you are encouraged to investigate and use them.

Gap reduction with a variety of measures.

The gap-reduction design can be implemented using test scores, classroom grades, attendance figures, and a variety of other data. Using test scores, however, involves some additional computational procedures for two important reasons:

- scores on different tests are non-comparable.
- test score scales, unlike other scales (e.g., number of days absent) do not have equal intervals.

Both of these problems must be dealt with mathematically to provide interpretable evaluation results.

Assessing Gap Reduction Using Non-Test Data

An example using attendance data.

Before going on to the more complicated issues, it will be useful to work through an example of gap-reduction analysis using attendance data.

Suppose that, during their first two months of participation, project students averaged 8.3 days absent. During the same period, comparison group students were absent 3.8 days. A year later, during the same two-month interval, project students were absent 5.2 days and comparison group students 3.9.

	Pre	Post
Project Group	8.3	5.2
Comparison Group	3.8	3.9

The *pretest gap* was 4.5 days (8.3 - 3.8).
The *posttest gap* was 1.3 days (5.2 - 3.9).
The *gap-reduction* was 3.2 days (4.5 - 1.3).

Amount vs. percent of gap reduction.

In this example, there is no real need to "massage" the data any further. It would be helpful, however, to express the amount of gap reduction as a percentage of the pretest gap. In this case 3.2 days is slightly more than 70 percent of 4.5 days. One could thus say that the project was successful in reducing the attendance-rate gap by approximately 70 percent after one year.

The *amount* of gap reduction and the *percent* of gap reduction may give quite different impressions as to how successful the project has been. It will, therefore, always be a good idea to present both in your evaluation report.

Examining other gaps.

Similar analyses could be carried out using classroom or report-card grades. This type of analysis is particularly appropriate for former LEP students who have graduated into mainstream classrooms. Whatever gap in grades may exist between the former LEPs and their non-LEP classmates shortly after mainstreaming should not increase with the passage of time.

If the gap remains constant, the former LEPs are at least holding their own. If it decreases, the former LEPs are continuing to catch up. But if it increases, the project failed to prepare its target students to progress effectively through the regular school program.

Assessing Gap Reduction Using Test Scores

Test scores may have unequal intervals.

Test scores, as mentioned earlier, lack some of the desirable features that other indices possess. A day absent is always a day absent--it makes no difference whether it is the second or the forty-second day absent or whether it occurred at the beginning of the project or a year later. A test score point, on the other hand, is likely to represent a different increment of performance at different performance levels, *and* at different times. Thus, pre- and post-test gaps that are equal in terms of test-score points may not be equal in terms of group performance levels.

Mathematical adjustments are appropriate.

These and other problems discussed below can be reduced to minimal proportions through mathematical manipulations. The required manipulations are performed automatically by the software that has been developed to accompany this *Users' Guide*, and you need not make any attempt to understand them unless you intend to do the analyses by hand or unless you find them intrinsically interesting. Manual procedures along with the rationales that underlie them are described in Appendix H.

Preparing data for analysis.

Four simple steps are required to prepare the data for either computer or manual analysis. First, you must be sure that the data you enter into the analyses are either raw (number correct) or scale scores[2]. You may not use percentiles, stanines, NCES, or grade-equivalent scores.

Scale scores are preferred to raw scores. Be sure to ask for them if you use a test-scoring service. If you are doing manual test scoring, however, it is probably not worth converting individual-student raw scores to scale scores.

All gap reduction calculations can be done using raw scores *if the treatment and comparison groups took the same level of the test.* If they were tested with different levels, you must work with scale scores--but you can convert mean raw scores to their scale-score equivalents. It is not necessary to convert individual-student scores.

Students with missing data are excluded.

Second, you must be sure that all students included in the analyses have both pre- and posttest scores. Students with only one or the other *must be excluded.* This rule applies to both project and comparison groups.

---

[2] Scale scores are given different names by different test publishers. These names include standard scores, expanded standard scores, Achievement Development Scale Scores, converted scores, and Growth-Value Scores. All of them refer to specially constructed scales that span the various levels of a particular test and thus provide a vehicle for converting out-of-level raw scores to in-level percentiles.

Part-year participants require special analysis.

Third, if any students were pretested late (because they entered the project after the regular pretesting had been completed) or posttested early (because they left the project before the regular posttesting) you must note the months when they were actually tested. If you use the computer program, it will ask for this information. If you undertake a manual analysis, you will have to use this information to calculate extrapolated test scores (see Appendix I).

Gap reduction with normative data.

Finally (but only if you use norms as your comparison group), you must find the raw or scale scores that correspond to the 16th, the 50th, and the 84th percentiles (for the appropriate grade levels) at both pre- and posttest times. This final data-preparation step is best explained by means of a concrete example.

Suppose you are working with a group of third-grade LEP students who were pretested in May of their second-grade year and posttested in May of their third-grade year. You must find the norms tables for the level(s) of the test you used. You must then find the raw- or scale-score-to-percentile conversion table that is appropriate for May of second grade (grade 2.8) and the one for May of third grade (grade 3.8).

Note: Test publishers will provide either a raw-score-to-percentile table or a scale-score-to-percentile table, not both. If you have a scale-score-to-percentile table, but want the raw scores corresponding to the 16th, 50th, and 84th percentiles, you will have to find the cor-

responding scale scores first and then convert them to raw scores using a separate raw-score-to-scale-score conversion table. You will have to follow a similar two-step process if you want scale scores but the publisher only provides a raw-score-to-percentile conversion table.

Using the computer program.

If you are using the computer program to do the gap-reduction calculations, you will simply enter the values you found in the norms table when the computer asks for them. If you are using manual procedures, step-by-step instructions are given in Appendix H.

In essence, the computer program and the Appendix H procedures "standardize" the four test-score data points (mean pre- and posttest scores for the project and comparison groups) so that scaling problems are minimized and scores from different tests are made comparable. It also calculates the amount of gap reduction from the standardized data points and generates another measure called the Relative Growth Index (RGI).

Understanding Relative Growth Indices (RGIs).

RGIs express, in percentage terms, the amount by which the progress of the project group exceeded or fell short of the progress of the comparison group. An RGI of +20% means that the progress of the project group was 20% larger than that of the comparison group. An RGI of -8% means that the progress of the project group was 8% less than the progress of the comparison group.

RGIs have one significant advantage over gap-reduction measures--they are independent of the heterogeneity of the comparison groups. What this means is that RGIs can be meaningfully compared between projects that use different types of comparison groups (notably grade mates versus norms). Gap-reduction measures will tend to be larger for grade-mate comparison groups than for norms because local groups tend to be less diverse than national samples--but RGIs will be unaffected.

When not to calculate RGIs.

RGIs also have one significant disadvantage when compared to gap-reduction measures. If the comparison group makes no progress or negative progress, the index becomes meaningless. RGIs should thus only be used in conjunction with achievement test scores (where posttest scores will almost certainly be higher than pretest scores for both the project and the comparison group). They should not be used when no "growth" is expected of the comparison group, as would be the case with indicators such as absenteeism rates or classroom grades.

When no growth is expected to occur in the comparison group, your analysis should end with gap-reduction calculations.

Using Medians Rather than Means

Medians may be used instead of means.

Means and medians are both statistics that summarize the performance levels of groups of students. If score distributions are symmetrical, the median will be approximately equal to the mean. Such equality is typical of the score distributions of norming samples on standardized achievement tests. Indeed, it is typical of the score distributions of most groups on well made tests that were neither too easy nor too difficult for them (see Chapter V).

Medians are preferable under some conditions.

In bilingual education projects, it is not uncommon for tests to be used that are too difficult for project-group students. When this occurs, the scores of those students are prevented from being as low as they should be. The result is that the mean of the group is artificially inflated.

While the group's mean score can be affected if there is even one student who does not know the answer to a single question, the group's median will not be affected unless there are many such students (theoretically up to half of those tested). Thus, if there is evidence that some students--but fewer than fifty percent of them--did not know the answers to any of the questions, the median provides an estimate of where the mean would have been had the test been able to measure lower performance levels accurately. Using the median in your gap-reduction calculations would thus be more appropriate than using the (spuriously inflated) mean.

Too difficult tests produce biased means.

The problem just described is most likely to be encountered by the project group at pretest time--and you should check that possibility carefully. The problem is less likely to occur at posttest time, but you should still check. It is also possible that the test may be *too easy* for the comparison group--particularly at posttest time. Whenever tests are too easy or too difficult, group medians should be used rather than group means. Otherwise, means are preferable to medians because they are more stable.

Software substitution of medians for means.

The computer program for this system automatically substitutes the median score of the treatment group for its mean score whenever the median score is one-fifth of a comparison-group standard deviation lower than the mean score. If you perform the gap-reduction computations manually, you should follow the same procedure.

It is quite acceptable to mix means and medians in a single analysis (although this practice would make it virtually impossible to determine the statistical significance of the findings).

Dealing with Regression Biases

Whenever a subgroup of individuals is selected from a larger group because they scored above or below some criterion (cutoff) value, their mean score on any subsequent testing will be closer to the mean of the original group than it was on the selection test. This relationship will hold regardless of whether the selected subgroup is retested immediately after their selection or at some later date. This phenomenon is called statistical regression, or regression to the mean.

Regression biases can contaminate growth estimates.

In bilingual education (where project participants are usually selected on the basis of low test scores), regression to the mean causes scores of the selected students to be higher on subsequent testings than they were on the selection test. Thus, if the selection test is also the pretest and the subsequent test is the posttest, there will be a pre-to-posttest gain that results solely from statistical regression. Care must be taken to recognize this gain as spurious and not mistakenly attribute it to the project.

Factors that affect the size of regression biases.

The amount of regression-effect bias that will occur when a set of test scores is used for both selection and pretest purposes depends on two factors: (a) the difference between the mean score of the selected subgroup and the mean score of the total group (the greater this distance, the more regression will occur) and (b) the correlation between the selection-pretest and the posttest (the lower the correlation, the more regression will occur).

When regression may be a problem

This type of pretest-based selection is likely to occur in many--perhaps most--Title VII projects. Consider the following scenario:

Students are selected for project participation on the basis of a home language survey and a subsequent English language proficiency test. Students scoring below a pre-established cutoff score are admitted to the project, while those scoring above the cutoff are not.

How to avoid regression biases.

The issue now becomes whether or not the scores obtained by the selected students will be used as their pretest scores for the evaluation. If not--in other words, if they are subsequently administered a separate pretest--there is no need to introduce a correction for the regression-effect bias.[2] If the scores students obtain on the English language proficiency test are also used as their pretest scores, a substantial amount of regression will occur between the selection-pretest and the posttest. You *must* correct for it.

---

[3] There *will* be a small amount of regression-effect bias even with a separate pretest. We regard it as too small to worry about, however, even though it could be statistically removed.

Before discussing the method of correcting for the regression-effect bias, it is useful to extend the scenario begun above and examine another situation that requires correction for regression.

Another problematic situation.

After one year of participation in the bilingual project, students are posttested. Assuming an annual testing cycle, the posttest scores for the first year will also serve as the pretest scores for the students who remain in the project for a second year. Some students may have achieved the exit criterion, however, and may leave the program.

If the year 1 posttest scores are used in any way to determine who should be exited from the project, you again have a situation in which (year 2) pretest scores were used to select the students who will remain in the project. Regression to the mean is the inevitable consequence, and you *must* correct for it in order to avoid biased growth estimates. Of course, you always have the option of avoiding the problem by not using selection test scores as pretest measures *and* by not using posttest scores in any way when you make exit decisions.

Automatic correction for regression biases.

The software that accompanies this *Users' Guide* automatically corrects for regression biases. Manual procedures are described in Appendix J. You should note, however, that the correction procedure assumes that selection into the project is based solely on the test scores that are subsequently used as the pretest measures.

If selection-test scores are combined with other measures to form a composite pretest, the correction procedures will overcorrect. Similarly, the statistical adjustment will be appropriate only if posttest scores are the sole criterion for exiting students from the program. If teacher judgments, classroom grades, or other considerations affect the existing decision, the adjustment will be excessive.

Bracketing the regression bias.

If the regression-effect adjustment is going to be excessive for one of the reasons just discussed, we recommend that you calculate gap reductions and RGIs both with and without the adjustment (the computer program does this for you). The "correct" values will then be bracketed between values that will be somewhat too low (those with the adjustment) and values that will be somewhat too high (those without the adjustment). Depending on the relative importance of test scores and other factors in the entry/exit decisions, you may be able to estimate approximately where the true values will fall within the bracketed range.

If posttest scores from one year are used both to help decide which students will be exited and as pretest scores for the non-exited students, *the scores of the exited students will be needed for computing the regression-effect adjustment.* Identifying these scores thus becomes a fifth step in preparing your test data for analysis (see pp. 89-90).

Assuring Representativeness of the Data

Evaluation findings should apply to all students served.

The current evaluation regulations specify that the findings of your evaluation must apply to all students served by the project. This requirement means that, if your project experiences high student turnover, you will need to make some provision for pretesting students who enter after the "regular" pretesting has been completed and/or posttesting students who leave the project before the "regular" posttesting is begun. Any student who has participated in the project for a minimum of 100 days should be considered as having been served, and all such students should be included in the evaluation.

Test scores of part-year participants require extrapolation.

Students who were served but who participated in the project for less than the full year will, presumably, have made less progress than full-year participants. Thus pooling the data from part- and full-year participants will tend to make the project appear less effective than it really was *unless* the data from the part-year participants are extrapolated to provide estimates of what their gains would have been with full-year participation. The computer software designed to accompany this *Users' Guide* performs such extrapolations automatically.

Manual procedures are described in Appendix I.

Separate analyses of full- and part-year students are advisable.

Even when the data from part-year participants are extrapolated to yield full-year estimates, you should be aware that growth may not be constant over the year so that part-year participants may have grown faster or more slowly than full-year participants--depending on which part of the year they attended. Also, the hierarchical nature of some learning tasks may make it difficult for late enterers to benefit from the project at all.

The preceding paragraph suggests several reasons why combining even the extrapolated data from part-time participants with those for full-time participants may yield misleading results. What we suggest is that you analyze the data several ways. We recommend, for example, that you calculate separate RGIs for part-year and full-year participants.

Still finer subgroupings may be desirable.

If you have enough part-year participants, you might also conduct separate analyses for late enterers and early departers--or for shorter-term and longer-term participants. You should not place too much confidence in analyses based on fewer than about 10 to 15 students, however, since the RGIs of small groups will be unstable. Given that general guideline, experiment with whatever subgroupings you think might make sense.

Going Beyond the Gap-Reduction Design

The regulations do not require that you go beyond the gap-reduction design. But the gap-reduction design does not provide a quantitative estimate of the project's impact. Such an estimate can only be obtained through proper implementation of an experimental or quasi-experimental design.

Many projects may not be able to implement an experimental or quasi-experimental design because the kind of non-project comparison groups they require may be excessively difficult or impossible to obtain. On the other hand, the value of your evaluation would be enhanced if you were able to obtain a sound estimate of project impact.

With this objective in mind, we have described, in Appendix G, three quasi-experimental designs that you *may* be able to implement--depending on your particular situation. We encourage you to read Appendix G and, if you can, to implement one of the designs presented there.

VIII. PROCESSING AND ANALYZING DATA

Two types of data must be dealt with: outcome and process.

There are two basic types of data to be reduced and analyzed: outcome data and process data. Generally speaking, outcome data are linked to individual students while process data are linked to various components or aspects of the project. In this chapter, we discuss data reduction and analysis for individual outcome and process measures. In Chapter IX we discuss integrating and interpreting the analytic findings.

Analyzing Student Outcomes

Outcome data include test scores and non-test data.

Student outcome measures include both test scores and non-test data (e.g., number of days absent). We begin with the test data.

Test scoring can be done either manually or by machine. We recommend machine scoring if you can afford it. The advantages pertain mostly to accuracy and can be sizeable, particularly if the scoring task includes the conversion of raw scores to scale scores.

Use a scoring service if you have large numbers of students.

Determining raw (number correct) scores by hand may not be excessively burdensome with small numbers of students. With large numbers of students, boredom, carelessness, and a desire to be done with the task often combine to produce high error rates. Quality control procedures are definitely in order.

Check for accuracy of manual scoring.

For each scorer, you should have a second person check the accuracy of a randomly selected 5% of the scores (e.g., for every 20 tests, rescore one). If any error is found, another 5% of the scores should be checked. If an error is found in this group, all the tests should be rescored.

Use scale scores with functional level testing.

As mentioned earlier, you *must* use scale scores if you do functional level testing. Without them you cannot quantify the gaps between groups that were administered different levels of the test.

Convert the mean raw score to its scale-score equivalent.

Unfortunately, converting raw scores to scale scores is an even more error-prone procedure than simple scoring when done manually. If you must use manual procedures, we recommend that you calculate mean raw scores for the various groups of interest and convert them to their scale score equivalents rather than converting individual-student raw scores.

You will not get exactly the same result with the short-cut procedure as with the more correct but cumbersome and error-prone alternative, but it will usually be very close. With machine scoring you should always request that the scoring service provide you with individual-student scale scores.

Clean answer sheets before machine scoring.

Whether you use machine or manual scoring procedures, you will have to do some preparation of the test booklets or answer sheets. In both cases you need to verify that the student identification information is compete and accurate. In the case of machine-scored answer sheets, you need to erase stray marks and make sure that each student's answers are marked clearly enough to be picked up. You will have to watch for other problems such as multiple answers to the same question, pattern responding, etc., if you do manual scoring. Machine scoring will pick them up automatically.

Compiling non-test data should be a relatively straightforward matter, but it makes a substantial difference whether the individual student data you need will be stored in district or project files and whether those files will be computerized or kept on paper. If the files are computerized, you will be able to do substantially more analytic work with them than if they are kept on paper--particularly if you can download data from the files directly into analytic routines. Sorting paper files and entering data manually can be very time-consuming activities.

Set up files for each student.

Regardless of whether data storage is computerized or not, you must set up files for each student in your project group (and possibly for each student in your comparison group). Each student's file should be labeled with his or her name and identification number and should include at least the minimal student characteristics information specified in Chapter IV of this *Users' Guide*. In addition, all test scores used for evaluation should be included in raw-and/or scale-score form along with the testing dates and the names, forms, and levels of the instruments used.

Include required non-test data in student files.

Each student file should also include number of days absent (broken down into at least two or three calendar periods), grade retentions (if any), disciplinary actions (if any), project entry and (if exited) exit dates, and referral to or placement in special education or gifted and talented programs. If the students have dropped out or entered postsecondary education institutions, those facts (along with their dates of occurrence) should also be included. Finally, it will be useful to include codes for the classrooms/teachers where or from whom students received project services. This information will be particularly useful if your data base is computerized.

Steps for manual data entry.

Coding and/or data entry. If data will be analyzed manually, you should:

- copy individual student data onto summary forms and sort the summary forms by class and grade level. (Remember, the correction for regression requires the posttest scores of last year's participants who were exited.)
- randomly check 5% of the entries as described above.
- make a copy of each summary form and store the copies in a location different from where the originals are kept.

Steps for computer-managed files.

For computer processing, data summary forms can also be developed to make data entry easier. However, if the data are well organized and you have a data management program available (dBase III, for example), you should enter data directly into the computer to save time and to reduce errors.

To check data entry, either

- randomly check what has been entered against the original records, or

- enter a sample of the data a second time and use a computer program to find differences between the two sets of records.

Constructing a data base. If you will be using manual data processing, each student's data (e.g., answer sheets and/or questionnaires) should be kept in a separate folder and filed so as to be easily retrievable.

Questions about constructing and using a computerized data base.

If you use a computer to create your data base, you must answer several questions:

- Will you use the "central" district computer or your own project or group computer?

- If you are going to use the central district computer, how are you going to add the unique project data to the district's centralized data base to create the project data base?

- If you are going to use your own computer to build your data base, will you be able to transfer some data from the centralized data base? If so, how?

Centralized data bases may meet most of your information needs.

Many districts keep centralized data bases on computers. Such records often include information on students' backgrounds, grades, and test scores. Where such a data base exists, it may represent a single source for much of the data you need.

If the district does have a data base, you should contact the person in charge of data processing for the district and find out if it would be possible to add your project data to the district data base. You should also explain the types of analyses you will be wanting to perform on the data.

Transferring data between computers.

If you decide to keep your project data on your own computer, but you also want to use the district's data base, careful planning can save you a lot of time and energy. First, you should find out:

- whether your computer and the district's computer can communicate with each other.
- whether the software on either computer will allow you to merge data files.
- If both of these operations are possible, you can simply transfer the district's data to your computer and merge the data, or if necessary, transfer your data to the district computer, merge data, and transfer everything back.

If neither of these operations is possible, then you may have to obtain printed copies of data from the district's computer and enter it into your computer file.

Six sound practices for data base management.

In setting up a data base, it is a good practice to:

- Store data in different subfiles. For example, there might be a student background data file and separate data files for each project year. If a student has the same ID number in each file, files can be merged as necessary for analysis.
- Ensure the confidentiality of the data by storing them in a computer file under the secret access codes which are available for some data management programs.
- Store multiple copies of files using different media (e.g., disk, tape, and hard copies).
- Examine files periodically, especially when new data are added, and make any changes to enhance their accessibility and safeguard against their misuse.
- Use interactive data management software whenever possible.
- Always include the district student ID number with each student's data.

The gap-reduction design meets all requirements for assessing academic progress.

Analyzing the data. The regulations require that you evaluate the academic progress of (a) LEP project participants, (b) non-LEP project participants, and (c) former project participants now in mainstream classrooms. All three of these requirements can be met using the gap-reduction design (see Chapter VII).

The regulations also require that, as appropriate, changes in student grade retention, dropout rates, absenteeism, referral to or placement in special education classes, placement in programs for the gifted and talented, and enrollment in postsecondary education institutions be assessed and reported.

Measuring changes in low-frequency events may require multi-year data.

The baselines from which to measure change are presumably the pre-project rates. Unfortunately, some of the rates, such as placement in programs for the gifted and talented, are likely to be quite low. For this reason, it will probably not be possible to obtain reliable measures of change from a single year's data. You will have to search historical files from several years to determine pre-project rates and then compile data for several project years before any changes will be detectable.

Pre-project rates may be included in the needs assessment.

A good needs assessment may have picked up on problems such as grade retentions and absenteeism. If statistics were compiled on those events and were included in the project application, they provide good baseline (pre-project) data. Statistics of the same type collected after a year or more of project operation should enable you to assess any changes that may have occurred.

Population shifts may invalidate before-after comparisons.

One problem you are likely to encounter is that school or district records may be inadequate or very difficult to access. A second problem is that the school population may have changed (perhaps due to an influx of refugee students or the creation of a new labor market). Under such circumstances, systematic differences may exist between the pre- and post-intervention samples of students that would invalidate any attempts to assess project impact. You should then present the data *and* explain why whatever change was observed cannot be attributed to the project.

Simple tabulations can be used to summarize project-related changes.

Under more favorable conditions, we recommend compiling data on an annual or multiple-year basis for periods before and during project operation. The analysis, then, need be no more than a simple tabulation. You might, for example, report, "163

(19%) of the 858 LEP children enrolled in school were retained in grade during the three years immediately prior to the start of the project. During the first three years of the project operation, that total dropped to 101 (12%) of the 842 LEP students enrolled in school."

If the sample size is sufficiently large (30 or more in each subgroup), you may want to analyze the trends by grade, by ethnicity, or by other variables of interest.

Contrast gains on curriculum-relevant and non-relevant items.

Secondary analyses can also be done to shed additional light on the findings of the primary (e.g., gap-reduction) analyses. Assuming that you classified test items as curriculum-relevant and non-curriculum-relevant as part of the test selection process, you could have the two sets of items scored separately. You would then expect to see a larger percentage increase (from pre- to posttest) for the curriculum-relevant items than for the non-relevant items.

Such a finding would tend to support the hypothesis of a positive project impact, while its opposite would suggest a negative impact. Evidence of this type can never be conclusive, however, since the non-relevant items might also be intrinsically more difficult than the relevant items or vice versa.

Classroom performance data can supplement other analyses.

Analyses of classroom performance (curriculum-specific objectives) as reflected by average grades, curriculum units completed, numbers of objectives mastered, etc., although not required by the regulations, could also be carried out. Such analyses might be particularly useful for identifying weaknesses in the project that need to be remedied (see below). They are also useful in confirming student outcome assessments derived from annual measurement of achievement growth.

Classroom grades and/or scores on teacher-made tests are likely to be available in most instances. If they are, it would be informative to calculate the correlation between classroom grades and posttest scores on your achievement test. (Your regional Evaluation Assistance Center can help you design such analyses). A positive correlation would provide *direct* confirmation that the test measured the same achievement dimension as classroom grades and *indirect* evidence that the test measured what was taught.

District or statewide testing programs can also provide useful data.

Scores from district- or statewide testing programs are likely to have more desirable characteristics than classroom grades, *if the tests are not too difficult for project students.* They could also provide useful confirmatory evidence if correlated with posttest scores from the project's evaluation test. Furthermore, if the district or state tests are administered on approximately the same schedule as the evaluation tests, you could correlate growth scores in addition to postproject measures. Several such analyses may be possible if you have an efficient, computerized database. Otherwise you may have to limit your analyses to those required by the regulations.

Analyzing Project Processes

Discrepancies are the key to process evaluation.

Process analysis is largely a matter of looking for discrepancies between the project *as intended* and the project *as implemented.* In some cases it may also be possible to quantify the magnitude and presumed importance of the observed discrepancies. Ultimately, the goal is to relate process discrepancies to possible discrepancies between anticipated and observed student outcomes for the purpose of effecting project improvement (see Chapter IX).

Process discrepancies can affect the total project (e.g., instructional materials arriving three months late) or individual classrooms (e.g., two of the project teachers speaking English with a heavy accent).

Process evaluation is not required by the regulations.

The regulations do not require any form of process analysis (the topic of the remainder of this chapter) nor any analyses that examine relationships between processes and outcomes (the subject of Chapter IX). The main advantages to be gained from evaluation, however, relate to project improvement. For this reason we urge you to undertake at least some of the analyses described in the remainder of this chapter and in Chapter IX.

Discrepancies need to be quantified.

All forms of process evaluation will benefit from quantification of the discrepancies between what was intended and what actually happened--but the need is greater and the task more difficult when there are multiple project classrooms per grade. In this situation, some sort of scale must be developed to reflect the differences that exist among them.

Discrepancies need to be quantified.

Scaling is no problem for discrepancies such as percent instructional usage of L1, or ratio of positively to negatively reinforcing teacher utterances. These discrepancies are basically self-scaling. Others, such as heaviness of teacher accent or teacher expectations regarding student achievement levels, however, can be difficult to quantify.

Avoid vague constructs. Instead, count observable events.

The amount of difficulty you are likely to experience in quantifying discrepancies will be directly related to the vagueness of the relevant constructs. We have already suggested (see Chapter IV) that complex constructs require high-inference observations and that replacing them with more directly observable (lower inference) substitutes will yield significant interrater-reliability benefits. Parallel benefits will accrue with respect to the quantification of discrepancies. Replacing "heaviness of accent" with "number of words mispronounced" (wrong sound or wrong syllable emphasized), for example, changes a high-inference rating task into a low-inference counting task. Not only will interrater reliability increase--so will the adequacy of quantification.

Rank teachers/classrooms rather than rate them.

If there appears to be no adequate, directly observable substitute for a complex conceptual construct (e.g., level of teacher expectations), our next recommendation is rank-ordering classrooms/teachers on the characteristics of interest rather than rating them. While ranking ignores the size of differences between the entities being ranked, rating tends to distort them. And if you use multiple observers (which you should do anyway to assess the reliability of your data) average rankings will recapture some information relative to the size of differences.

Try to make your data collection non-threatening.

Either rating or ranking teachers can be problematic due to teacher-union rules or other political constraints. This is another reason for attempting to make your process data as objective as possible. Counting frequencies of behaviors is less threatening than evaluating teacher performance on some dimensions. In any case, you need to make it clear that you are *documenting project processes,* not *evaluating teacher performance.* Then, even if you are denied the opportunity to conduct classroom observations, you may be able to obtain reasonably accurate data by means of teacher interviews or questionnaires.

Once you have quantified discrepancies for classrooms/teachers, you can calculate mean or median values for each grade level served by the project. You will want to have classroom-level data for within-year analyses and project-level data for between-year analyses. This distinction and how the different types of analyses can inform you about process-outcome relationships will be the subjects of the next chapter. Before looking at these causal relationships, however, it may be useful to examine process data in somewhat greater depth.

Combine discrepancy scores to create a project implementation index.

There are several things you could do that would simplify analyses of process-outcome relationships and perhaps make them easier to interpret. First, you could combine discrepancy scores across processes and thus obtain an overall project-implementation score for each classroom.

A simple, but often less than ideal, approach is simply to sum the individual discrepancy scores. The problem with this approach is that it will automatically weight each component of the composite score according to its variability across projects. Thus, if there are large differences among classrooms with respect to a particular process, that process will be more heavily weighted in the composite scores than a process that is less variable across classrooms.

Standardize discrepancy scores to weight them equally.

The recommended procedure for dealing with this problem is to calculate the standard deviation of discrepancy scores for each process across all classrooms and then divide each individual score (for the same process) by that standard deviation. Summarizing the "standardized" discrepancy scores for each classroom will then give you composite implementation scores in which each process is approximately equally weighted.

You may stop at this point, or you may wish to weight the scores according to your assessment of their importance. If you wanted to give twice as much weight to instructional usage of L1 as to pronunciation of English words, for example, you would simply multiply each classroom's standardized discrepancy score for instructional usage of L1 by 2. To obtain the properly weighted composite score, you would then sum the weighted, standardized discrepancy scores for each classroom.

Block teachers/classrooms into analytic groups.

A second procedure that will be useful for basic analyses of process-outcome relationships is called "blocking." Instead of working with individual classroom scores, you could sort (block) the classrooms into high, medium, and low groups.

The blocking process will serve the purpose of increasing the stability of your data. The mean score of a group of classrooms will be more stable than the scores of the individual classrooms. The increased stability will have the effect of making your analyses more sensitive to *real* relationships between process and outcome variables. This topic is discussed more fully in Chapter IX.

IX. INTEGRATING AND INTERPRETING RESULTS

Begin by examining your outcome results.

If you have carried out all the procedures recommended in the previous chapters of the *Users' Guide*, you will have collected and analyzed two types of data: process data and outcome data. Although you will have looked at outcome test results for the project as a whole, you need to examine them carefully by grade level and possibly by classroom. Deciding whether they exceed, meet, or fall short of your expectations should be the first step in data interpretation.

Interpreting Gap Reductions and RGIs

Disappointing outcomes need an explanation.

Perhaps you are perfectly satisfied with your outcomes, but often some or all of them will be disappointing. If your gap reductions and RGIs are lower than you expected them to be, you will want to find out why. You can address this question by proceeding through the steps outlined in this chapter.

Positive outcomes also require interpretation.

Even if your gap reductions and RGIs are better than expected, you should not skip the step of data interpretation. You need to integrate and interpret your data in order to establish causal linkages between project processes and student outcomes. High RGIs may not have resulted from the project.

Consider the possibility that they are high because students became much more test wise between the pretest and posttest, for example. If this were indeed the case, you should expect smaller RGIs in following years, even if the project is working as planned. As another example, suppose a high RGI is due to comparison group students being post-tested with too easy an instrument. If in a following year, a higher level of the test is used, project outcomes will appear to be less satisfactory. Thus, it is never enough to say that your project met its outcome goals. You must also be able to suggest reasons for its success, and to show evidence that it was the project, and not some unrelated events, that produced the outcomes.

Compare RGIs to other outcome data.

Integrating test data with other outcome data. The first step in interpreting RGIs is to look at the other outcome data you have collected. These may be data on absenteeism, scores on teacher-developed tests, or mainstream classroom grades of former participants. Do these additional data seem to confirm the RGIs, or call them into question? If all your outcome data point in the same direction, you can dismiss some of your doubts about the validity of your test score findings. If outcome data conflict, however, further investigation is necessary.

Conflicting outcome data call for detective work.

If there are conflicts between other outcome data and standardized achievement test scores, consider the possibility that the standardized test is invalid for your project. If you did not go through the test selection steps presented in Chapter V of the *Users' Guide*, do so now. Your RGIs may be a function of poor test selection and may not reflect project impact.

Look for patterns in outcome data.

Perhaps only some of your other outcome data conflict with your RGIs. For example, suppose your RGIs in English language arts are low, but at some grade levels students' scores on published curriculum-based tests have been quite satisfactory. This pattern of results suggests both that the content and validity of the achievement test was low (it was insensitive to performance differences that the curriculum-based tests picked up) and that the language arts component of the project was more effective at some grade levels than at others.

If absenteeism continued to be high at some grade levels, but improved at others, it might be logical to infer that affective components of the project were better implemented at the latter grade levels than at the former.

Organizing data by student groupings.

You should always group your data by grade level. Then ask, "are some grade levels consistently lower or higher than others?" You should also group data by classrooms or groups of classrooms and ask, "are there noticeable differences between classrooms?" If your students are heterogeneous, you might try grouping them by factors such as home language or parent literacy. Perhaps your project has been successful with some groups and not with others.

Always consider the possibility that your data may be invalid.

If your outcome data are mixed and show no discernible pattern, you are in the unfortunate position of having few conclusions to draw about project outcomes. Under such circumstances, you should re-evaluate the validity of all your outcome measures by reviewing the following questions:

- Was the test used to obtain RGIs valid and appropriate for the project and comparison groups?
- Were all outcome data collected in a careful, consistent manner? Were there irregularities in test administrations?
- Were computational errors possible in data analysis?
- Were test scoring errors possible?
- Did you use the correct conversion and norms tables in your computations?

What caused student outcomes?

Integrating outcome data with process data. To interpret your outcome data, you need to identify the most likely cause of the outcomes you obtained. If you obtained high RGIs, but had a minimal treatment (e.g., 10 minutes of ESL per week), you can assume that the project did not cause the high RGIs. You should try to find out what did. If you had a well implemented, comprehensive treatment but low RGI's, you should also try to determine the reason. Whatever your outcome data look like, you owe it to everyone involved in the project to try to determine:

- what project features, if any, most likely contributed to the outcomes.

- what project features did not appear to affect the outcomes.

- what nonproject features played a role in the outcomes.

You should address these questions by examining your process data together with your outcome data.

Review your process data for surprises.

As a first step, look over all your process data--teacher, project, and student characteristics; implementation data; and non-project events. Are there any unexpected elements in the data? For example:

- Were teachers less qualified than the project design called for?

- Did the project serve students from unexpected ethnic or language groups?

- Did the project serve more (or fewer) students than expected?
- Did classroom observations reveal failures in implementation or changes to the project design?
- Was the project allocated necessary space and equipment?
- Was parent involvement as extensive as planned?
- Were necessary materials available?
- Did teachers attend planned preservice and in-service sessions?
- Could other specific events in the school or community have affected student outcomes?

Higher implementation should lead to higher outcomes.

Once you have identified areas in which the project did not operate as planned, you should review related outcome data. For example, if second-grade teachers received the lowest implementation scores of any grade level observed, you would expect to see poorer project outcomes at that grade level. Conversely, if fifth grade teachers had the highest implementation scores, you would expect higher outcome scores at that grade level. If both situations pertained, you would certainly expect fifth grade outcomes to be better than second grade outcomes. If this is the case, it suggests that when the project is properly implemented, students achieve more.

On the other hand, process discrepancies may not reduce project impact and may even enhance it. Placing bilingual paraprofessionals in classrooms taught by teachers who are without bilingual credentials *could*, for example, prove to be a more effective strategy than the originally intended use of credentialed bilingual teachers without aides.

Project emphasis should *be reflected in outcomes, but surprises occur.*

Sometimes your project may have unexpected outcomes. Suppose that your project emphasized basic skills instruction heavily, but was not at all prescriptive about how science and social studies should be taught. Your outcome data show that students made a very impressive jump in social studies--a much better outcome than in reading or math. You should consider the following questions:

- If the social studies test is in English, did it serve as an English reading test? In that case, students' scores may reflect increased English proficiency.
- Did the social studies test have higher content validity than the other tests?

- How was social studies taught? Did teachers use techniques that should be considered for other instruction?

Patterns of outcome and process data.

Where process data show a significant discrepancy from the project design, you would expect to see lower outcomes. This could happen in specific classrooms, schools, subject areas, grade levels, language groups, or years. You will need to group students in ways suggested by your process data and calculate RGIs to see what the pattern is. An effective project should produce lower RGIs at points of known failure or weakness. The pattern will not necessarily be completely consistent, but it should be suggestive.

Display your data graphically.

The simplest way of examining patterns is to make a graphic display of your data. As an example, suppose that you have overall implementation scores for every classroom in your project and you also have classroom-level RGIs. You could use a bar graph such as that shown in Figure 4 to illustrate the relationship (which in this instance is quite strongly positive).

The situation can easily be more complicated, however. Some project components may not be effective with your project group. Failures to implement them properly may not show up as reduced outcomes. Whatever substituted for proper implementation may even have produced better outcomes. You need to be alert to this type of unexpected relationship, so that you can avoid taking inappropriate corrective actions.

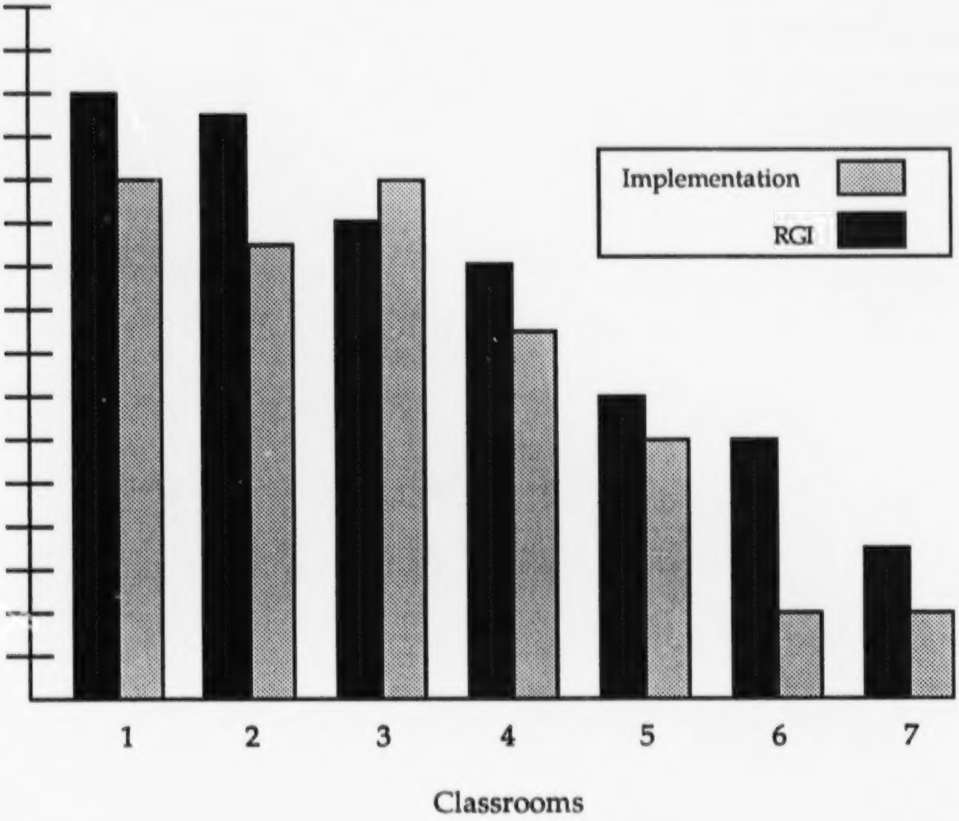


Figure 4. Graphic display of outcome and process data.

There may be no patterns.

Perhaps there is no discernible relationship between process and outcome data. Consider these questions:

- Are the process and outcome data valid?
- Was the project as implemented substantial enough that one could reasonable expect it to make a difference?
- Were student needs accurately defined?
- Was the project designed to meet those needs?
- Were the outcome measures designed to measure whether those needs were met?

Drawing Conclusions and Making Recommendations

The ultimate purposes of your project evaluations are to enable you to draw conclusions about project impact and to make recommendations about project improvement. Your conclusions and recommendations should derive from your integration of process and outcome data. Outcome data alone, do not demonstrate a specific cause. Causal relationships are only credible when the linkages between processes and outcomes both make sense and are supported by the data.

Show evidence that the project caused the outcomes.

In drawing conclusions, begin with any positive outcomes. Can you show evidence that the project contributed to these outcomes? If not, is there evidence to suggest other causes? In the light of these findings, what should the project do next year (i.e., modify the project, modify the evaluation, or continue as planned).

Next proceed to disappointing outcomes. Do these appear to be related to specific project weaknesses? If so, how can the project be strengthened? Are the outcomes possibly related to problems in the evaluation? Can these be corrected in succeeding years? If the results are uninterpretable, how can they be made more interpretable in the future?

Were your expectations too high or too low?

Finally, look at your expectations for outcomes in the light of what you now know about your project and your students. Were your expectations too high? If your project is necessarily a very limited one, your results will probably be limited also. High student mobility, for example, will dilute the effects of your project, and there is not much that can be done about it.

Perhaps your expectations were too low. If the project met or exceeded all your expectations, it may be appropriate to raise your sights in following years.

Confer with your team before finalizing your recommendations.

Conclusions and recommendations should be discussed with your evaluation team, if at all possible. They may have insight into project weaknesses and/or sources of inaccuracies in evaluation data.

X. REPORTING

Your report must address all points specified in the regulations.

The regulations require that evaluation reports include specific information about the students served, the services provided, the project staff, and the impact of the project on the targeted students. All of these requirements have been discussed at appropriate places throughout this *Users' Guide*. To refresh your memory we also provide a checklist of those information items later in this chapter.

Prepare your report to satisfy all your audiences.

Obviously your evaluation reports must contain all of the information required by the regulations. Presumably, however, the Federal government will not be your only audience. You will have to cover the information needs of your other audiences as well--unless, of course, you are able to prepare separate reports for different audiences.

Since your audiences are likely to range all the way from policy makers (school board members, school and district administrators), who will have no interest in details, to educational researchers who will be concerned with the soundness of methodologies you employed and the credibility of the results you obtained, the organization and presentation of your report deserve careful consideration.

Follow a few basic reporting guidelines.

We recommend that you begin your report with some sort of non-technical summary for administrative personnel. Details describing the implementation of the evaluation and the processing of data should be presented later (perhaps in technical appendices) to satisfy researchers that all threats to validity were dealt with adequately and that the findings do not reflect large and/or unacknowledged biases.

In preparing the report, remember to avoid technical terms as much as possible. Use the active voice, and don't be afraid to use "we." A visually appealing format and graphic presentation of data make a report easier to understand. Finally, when possible, make a verbal presentation of the results to interested audiences.

Executive Summary

Begin by listing major findings. Keep it short.

You should begin your "Executive Summary" with a *very* brief description of the project being evaluated, followed by a presentation of the *major* evaluation findings. Findings should be described in order of their importance (your readers may not get past the first few--especially if they seem unimportant) and as briefly as possible (if they are brief enough your audience might even read all of them).

Start with positive findings.

Although your discussion of findings will be limited to the major ones, it is desirable to mention that there *are* additional findings. You can then reference the location(s) in the main body of the report where they are discussed.

It is usually a good idea to present positive findings first, and if reasonable, to describe negative findings as "areas for project improvement" or some similar euphemism.

Include a brief summary of earlier years' findings.

If your report deals with the second or subsequent year of a project operation you should briefly describe the findings of earlier years, changes made to the project as a result of earlier evaluations, and any improvements in outcomes that may have resulted from those changes. No further information need be included in your summary, which will be most effective if you can keep its total length within five pages.

The Main Body of the Report

Your Introduction *should cover seven topics.*

The main body of the report should begin with an *Introduction* that includes the following information:

- A brief description of the community, district, and school settings (1 page).

- A description of student characteristics (1 or 2 pages).

- A presentation of student needs assessment findings (1 page).

- A narrative description of the treatment. This description should mention whatever differences there were between actual and intended treatment characteristics, but without excessive detail (3 pages).

- A brief description of staffing and staff qualifications (1/2 page).

- A brief description of materials and materials-development activities (1 page).

- A brief description of parent and community involvement activities (1-2 pages).

Cover four topics in the Methodology *chapter.*

The second chapter of the report should describe the evaluation's *Methodology.* It should:

- List the project's outcome objectives and describe how instruments and procedures were selected/developed to assess each of them.

- Describe how and by whom both process and outcome data were collected.

- Summarize the data processing and analysis activities that were undertaken.
- Indicate what steps were taken to avert threats to the evaluation's validity (in lay person's language--technical detail should be relegated to an appendix).

Organize findings by objective.

The next chapter should report the evaluation's *Findings*. This chapter should be organized by objective. Findings should be broken down by grade level within each objective and by student group (LEP participants, non-LEP participants, and exited former LEP students currently in mainstream classrooms) within grade.

A discussion of the quality of the data supporting each finding should be integrated into the presentation. Again, however, excessive detail should be relegated to an appendix.

End with a Conclusions and Recommendations *chapter.*

You may include a *Discussion* chapter if you feel it is warranted. Otherwise you may go directly to the report's final chapter, *Conclusions and Recommendations.* In this chapter you should attempt to integrate data from the *Findings* chapter to address such broad questions as:

- How well are project participants progressing academically?

- What evidence is there that they are benefiting from project services?
- Are the findings consistent with theoretical expectations?
- Are observed outcomes better or worse than expected?
- What project components appear to be most effective?
- What project components appear to be least effective?
- What changes could be made to the project that would (presumably) enhance its effectiveness?
- What differences are there between this year and last year?
- What changes might be expected next year?

Don't be afraid to speculate.

In most cases, your attempts to address these questions will require speculation. You should always indicate the degree of confidence you have in your hypotheses and the extent to which the data tend to support them and why. Do not hesitate to present alternative hypotheses along with the evidence supporting each.

Describe specific procedures for implementing your recommendations.

If possible, provide a step-by-step guide for each recommendation you make, suggesting how it should be implemented. Concrete, specific steps are more likely to be carried out than general suggestions.

Strive for objectivity. Excessively optimistic interpretations will cast doubts on the credibility of the entire report. On the other hand, informed speculation on your part can be helpful to the reader. Be sure to present the data that led to your speculation so your readers can draw their own conclusions.

Checklist of Mandated Reporting Requirements

Be sure your report covers the mandated topics.

The current (1986) regulations governing bilingual education evaluation are *explicit* about certain reporting requirements. These requirements include:

Characteristics of the students served

- educational background.
- assessed needs.

- academic competencies in English language proficiency, native or other language proficiency (for projects of developmental bilingual education), and other academic subjects.

Characteristics of the treatment

- specific educational activities undertaken.
- pedagogical materials, methods, and techniques used.
- relative amounts of instructional time spent with students on specific tasks.

Characteristics of the project staff

- professional qualifications.
- language competencies.

Student outcomes

The following information must be reported separately for (a) LEP project participants, (b) project participants whose primary language is English, and (c) former LEP participants currently in mainstream classrooms:

- amount of time students received instructional services in the project or in another instructional setting.

- progress (as reflected by test scores) students have made in English language proficiency, L1 proficiency (for projects of developmental bilingual education), and other academic subjects.

- changes in the rate of student (a) grade retention, (b) dropout, (c) absenteeism, (d) referral to or placement in special education programs, (e) placement in gifted and talented programs, and (f) enrollment in postsecondary educational institutions.

Also *implicit* in the regulations is the need to provide a description of the evaluation methodology, including particularly:

- steps that were taken to assure that the evaluation findings would be representative, i.e., could be generalized to the entire population served.

- steps that were taken to assure that the data collection instruments and procedures were reliable and valid.

- steps that were taken to assure that evaluation procedures minimized error (averted threats to validity).

These implicit requirements should be addressed in the *Methodology* chapter.

You should contact your local Evaluation Assistance Center if you need help with any of these reporting requirements.

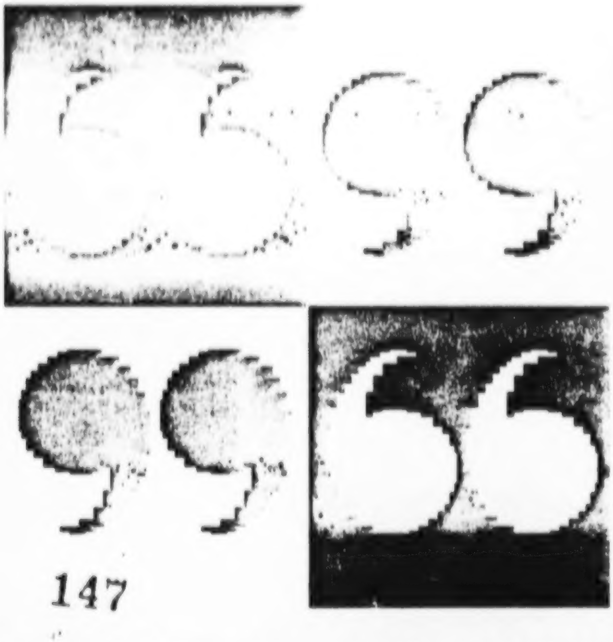
Good luck!

Bilingual Education Evaluation System

Users' Guide

Volume II - Technical Appendices

November 1987



RMC Research Corporation -- 2570 El Camino Real West, Mountain View, CA 94040

BILINGUAL EDUCATION EVALUATION SYSTEM

USERS' GUIDE

Volume II. Technical Appendices

G. Kasten Tallmadge
Tony C. M. Lam
Nona N. Gamel

November, 1987

Prepared for:

U.S. Department of Education
Washington, D.C.

By

RMC Research Corporation
2570 W. El Camino Real
Mountain View, CA 94040

The research performed herein was performed pursuant to Contract No. 300-85-0140 with the U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, represent official U.S. Department of Education position or policy.

149

TABLE OF CONTENTS

Page

List of Tables and Figures.. iii

Appendix A. Classroom Observation.. A-1

Appendix B. Interrater Reliability.. B-1

Appendix C. Functional Level Testing.. C-1

Appendix D. Translating Tests... D-1

Appendix E. Major Publishers of Nationally Standardized Achievement Tests.. E-1

Appendix F. Test Reliability... F-1

Appendix G. Quasi-Experimental Designs.................................... G-1

Appendix H. Gap Reduction Calculations..................................... H-1

Appendix I. Extrapolation Procedures.. I-1

Appendix J. Correcting for Regression.. J-1

LIST OF TABLES AND FIGURES

Page

Table A-1. Advantages and Disadvantages of Classroom Observation.......... A-2

Table B-1. Agreement/Disagreement Matrix of Observations.......... B-3

Table B-2. Ratings of Two Independent Observers.......... B-6

Table B-3. Data Required for Calculating a Rank Order Correlation.......... B-6

Table F-1. Percent Increase in the Standard Error of the Mean when a Less Reliable Test is Selected.......... F-8

Table G-1 Values of ± .05 as a Function of Sample Size (N).......... G-11

Table J-1. Table for Estimating the Total Group Pretest-Posttest Correlation.......... J-5

Figure A-1. Example of a coded behavior record.......... A-4

Figure C-1. Example of student scores on two vocabulary tests.......... C-3

Figure G-1. The regression-discontinuity design (in its simplest form) with a substantial treatment effect.......... G-18

Figure H-1. Illustration of gap reduction.......... H-5

APPENDIX A

CLASSROOM OBSERVATION

Classroom Observation

Observations can be conducted informally or formally. Informal observation is not guided by any predetermined scheme and the data or reports it produces are based primarily on subjective judgment. Formal observation is structured following specific rules for observing certain behaviors. Three generally recognized formal observation procedures--behavioral checklists, coded behavior records, and delayed report instruments--are discussed in this appendix.

Formal observational techniques are most appropriate for documenting instructional activities, verifying self-reports, and for identifying discrepancies between intended and actual program implementation. They can also be used to evaluate the extent to which the teachers apply in the classroom what they learn in staff development training. Table A-1 lists some of the advantages and disadvantages of formal classroom observation as an evaluation method.

Table A-1

Advantages and Disadvantages of Classroom Observation

ADVANTAGES	DISADVANTAGES
• Observation can be highly credible when seen as the report of what actually took place presented by disinterested outsider(s).	• The presence of observers may alter what takes place.
• Observers provide a point of view different from that of people most closely connected with the program.	• Time is needed to develop the instrument and train observers.
	• Credible observers must be located.
	• Time is needed to conduct sufficient numbers of observations.
	• Scheduling problems are common.
	• Observation data can be very difficult to analyze and interpret.

Behavioral checklists. Behavioral checklists usually include a list of a few (no more than 10) behaviors that should (or should not) occur in the classroom. The task of the observer is simply to make a tally mark on the checklist each time he or she observes the behavior taking place. Sample behaviors include:

- Teacher asks a question in L1.
- Student asks a question in L1.
- Student uses English to answer a question posed in English.
- Student uses English to answer a question posed in L1.
- Student uses L1 to answer a question posed in L1.
- Student uses L1 to answer a question posed in English.
- Teacher translates (into English) student's question posed in L1.
- Teacher reinforces student's English-language response using L1.

As should be apparent from this listing, an almost endless variety of behaviors could be included. Since observers cannot deal with more than 10, however, it is crucial that you identify the specific teacher and student behaviors that are thought to be most critical to the project's success.

By reviewing the project's description, plans for staff development, lists of materials, etc., you should be able to identify teacher and student behaviors that are relevant to its instructional intent. You should then prepare an exhaustive list of such behaviors (perhaps including descriptions of behaviors that are counter to the project's instructional intent and should *not* occur). Final selection of the behaviors to be included in the classroom observation instrument should be made by the project designer(s).

Coded behavior records. Coded behavior records are more sophisticated than behavioral checklists in two respects: (a) they allow for the recording of more behaviors, and (b) they make provision for the recording of sequences of behaviors. On the other hand, they have the disadvantage of being difficult to use and difficult to analyze. Unless you feel you really *must* know about sequences of behaviors we recommend that you not adopt coded behavior records.

Coded behavior records usually consist of no more than a series of boxes. Behaviors are coded sequentially into the boxes using predetermined symbols. You might, for example, use the designation TQ1 to represent the teacher asking a question in L1. Then, if the student to whom the question was addressed answered in English, you might enter SAE into the next box. Figure A-1 provides an example of a coded behavior record.

TLE	TQE	SAE	SQ1	TAE			

S = student
T = teacher
L = lecture
Q = question
A = answer
E = English
1 = L1

Figure A-1. Example of a coded behavior record.

A three-element code is common, but more complex codes are possible. You might, for example, decide that it is important to know what kind of questions (fact versus inference, perhaps) the teacher asked in addition to the language in which the question was posed.

When developing a coding system, remember that classroom events occur rapidly. Too complicated a coding system will overload the observer and result in lower quality data than a simpler system. Make your coding system as simple as possible and verify its feasibility by trying it out in actual classrooms.

If possible, have more than one observer use the instrument to record data in the same classroom at the same time. If their coded records are significantly different, the difficulty *may* by an excessively complicated coding system.

Delayed report instruments. The delayed report technique is the least formal of the three classroom observation techniques discussed here, and the easiest to implement. In it simplest form it involves no more than observing events and activities for a prescribed period of time (say, five minutes) and then writing down a description of what was observed. The recording instrument could be no more than a blank piece of paper.

The reliability of observational data can be increased by using somewhat more structured delayed report instruments. For example, the observer might be instructed to focus on two or three students who were working together and to record the percentage of a four-minute period they used English to discuss their task.

The delayed report instrument might then include a question like the following:

The small group of students used English about what percent of the time?

_____less than 25%
_____between 25% and 50%
_____between 50% and 75%
_____more than 75%

Observation time periods should be short. Observers should record their observations immediately after the end of the time period and should be watching for a small number of events or activities.

One advantage of the delayed report observation technique is that it enables the collecting of "richer" data than the two techniques described earlier. On the other hand, richer data might require higher levels of inference to be made by the observer.

Consider the following question:

How would you describe the classroom climate?

_____positive	_____neutral/negative
_____enthusiastic	_____apathetic
_____task-oriented	_____disorganized
_____all of the above	_____all of the above

Clearly the observer is being asked to make a subjective judgment, and different observers would be likely to respond differently unless they had been thoroughly trained to interpret symptoms in the same way. To construct a useful delayed report instrument, project staff can meet and develop a list of questions about what goes on in project classrooms. The questions should then be tried out in actual classrooms to determine how accurately and reliably they can be answered.

The training of observers who are subsequently called upon to make subjective judgments necessarily involves identifying behavioral manifestations of such constructs as "positive climate." An effective shortcut might therefore be to address your questions to the directly observable manifestations (e.g., teacher frequently smiles and uses the word good) which comprise the construct rather than to the construct itself.

Selecting and Training Observers

The essential qualifications of the observer(s) will depend upon the observation method you choose. The more structured the method, the less knowledge of bilingual education the observer will require. In all cases, the observer should be objective and a non-stakeholder.

After teaching the observer the observation procedure, you should give the observer an opportunity to try out what he or she has learned. This can be done with a non-project class or with a videotape recording of a real class operation.

During the "practice" observation or the first project class observation, the ratings of each observer should be compared with those of the trainer. The latter serve as the criteria. If a videotape is used, the consistency of the trainee's ratings can also be determined by comparing successive trials. Depending on the kinds of ratings generated, different coefficients can be computed to reflect the extent to which the trainee and the trainer agree (see Appendix B).

You should identify and discuss the similarities and differences in the observation data produced by each rater and the trainer. If the discrepancies are large, further training or a new observer may be needed.

Growth or change estimates derived from observational data are likely to be biased if the collectors of such data *become more proficient* (possibly as a result of practice) or *less proficient* (possibly as a result of boredom or forgetting) over time. To minimize this source of potential bias, it is a good idea to provide some refresher training before each round of data collection. The need for consistency should be emphasized, and data collectors should be urged to follow documented procedures--even if they see room for improvement.

The difficulties associated with collecting valid and useful observational data do not end with instrument development or observer training. An additional difficulty is that students and teachers must get used to the presence of an observer before they will behave normally in front of one. You must be sure that this type of desensitizing has taken place before you begin "real" data collection.

To be sure that the classroom events and activities you are observing are normal, you should observe each classroom on a number of different occasions and avoid conducting observations when atypical events such as parties or special performances are scheduled. Too small a number of observations will be likely to yield non-representative and unreliable data. Spreading your observations out over the entire school year will also enable you to detect changes that may suggest improved implementation as the project matures or worsening implementation as teachers "slip back" into earlier behavior patterns.

If limited resources preclude your observing all project classrooms, you should select a sample that represents the full range of student, setting, and treatment characteristics. It will be more cost-effective to observe a sample of classrooms on multiple occasions than all classrooms only once or twice.

General Principles Applicable to Classroom Observation

- Make the observation instrument simple to use.
- Conduct multiple observations, spread over the entire year.
- Observe different features in different sessions.
- Observe a representative sample of classrooms on multiple occasions rather than all classrooms only once or twice.
- If possible, check intra- and/or interrater reliability.
- Avoid conducting observations when atypical events are happening.
- If possible, avoid informing teachers the exact date when the observations will occur.
- Try to conduct observations as unobtrusively as possible.

Where to Get Additional Information

The preceding discussion regarding classroom observation strategies and instruments has necessarily been somewhat superficial. Although the important points have all been covered briefly, you would be well advised to do some additional reading before actually attempting to collect observational data. The following sources are recommended:

Morris, L. L., & Fitz-Gibbon, C. T. (1978). *How to measure program implementation.* Beverly Hills, CA: Sage Publications.

Stallings, J. A. (1977). *Learning to look: a handbook on classroom observation and teaching models.* Belmont, CA: Wadsworth Publishing Company.

APPENDIX B

INTERRATER RELIABILITY

A persistent problem with human observers is that they can look at the same things but perceive them differently. Two different observers for example, might describe the same classroom in totally different terms.

Observer A	Observer B
Noisy, chaotic, and out-of-control	Individualized, spontaneous, and creative

The difference between these two descriptions is clearly not a function of the classroom itself. It results from different observer perceptions and is a classic example of low interrater reliability. If the two observers were asked to count the number of teacher-initiated positive and negative reinforcements of students' behavior during a given time period, we would probably find more agreement between their responses. Reliability, or lack thereof, may be just as much a function of the rating task as it is of the rater.

Usually, interrater reliability can be improved either by training the observer or, alternatively, by making the rating task more objective. All other things being equal, for example, you will obtain higher interrater reliabilities if the task is to count the frequency with which the teacher uses the word "good" than if it is to rate the supportiveness of the classroom climate. Even with extensive training, you may not be able to obtain acceptable levels of interrater reliability with the kind of high-inference tasks represented by describing the climate of a classroom.

Quantifying Interrater Reliability

In order to assess whether different observers would judge the same situation in the same way, you need to measure their reliability. Once you have quantified interrater reliability, you know whether to proceed with the observations or whether additional work will be needed before you can use your observers in your evaluation.

Different quantification techniques are required (and/or appropriate) for different rating tasks. For example, one of the examples used above requires observers to categorize teacher behavior, while another requires them to rate classroom climate. For different types of tasks, different methods are appropriate for determining interrater reliability.

Categorizing. Several techniques are available for assessing reliability in categorizing tasks. One of the simplest and best, however, is coefficient Kappa.

Suppose that Observer A and Observer B independently rate 200 teacher utterances as to whether they are positively reinforcing (P.R.), neutral (N), or negatively reinforcing (N.R.). Their ratings could be tabulated as shown in Table B-1 below.

Table B-1
Agreement/Disagreement Matrix of Observations

		Observer A			
		P.R.	N.	N.R.	Total
Observer B	P.R.	66	8	6	80
	N.	6	35	9	50
	N.R	7	14	49	70
Total		79	47	64	200

The number 66 in the upper left hand corner of the matrix shown in Table B-1 means that both observers agreed in categorizing 66 of the teacher utterances as positively reinforcing. The number 7 in the lower left hand corner means that 7 of the teacher utterances rated as positively reinforcing by Observer A were judged to be negatively reinforcing by Observer B.

The values in the main diagonal (the diagonal that runs from the upper left to the lower right of the matrix and, in this case, includes the numbers 66, 35, and

49) represent agreements between the two observers. The off-diagonal entries represent disagreements. In calculating Kappa, we only need to be concerned with cells on the main diagonal -- those representing agreements.

Sometimes interrater reliability is incorrectly expressed simply as the percent of the total number of observations that are agreements. Agreements will occur by chance, however, even if the ratings of the two observers are totally unrelated. Kappa (appropriately) corrects for chance agreements.

The number of agreements that would be expected to occur by chance in any cell of the main diagonal is given by the corresponding row total multiplied by the corresponding column total and divided by the grand total. The chance expectations for the middle cell of the matrix shown in Table B-1 is thus (47 x 50) ÷ 200 or 11.75. It is because the actual number of agreements (35) is greater than the chance expectations that interrater reliability is greater than zero.

The actual calculation of Kappa proceeds as follows:

Step 1. Sum the actual number of agreements (the values in the main diagonal of the matrix).

Step 2. Calculate and sum the chance expectations for each cell in the main diagonal.

Step 3. Subtract the result obtained in Step 2 from that obtained in Step 1.

Step 4. Subtract the result obtained in Step 2 from the total number of ratings.

Step 5. Divide the result obtained in Step 3 by the result obtained in Step 4. The "answer" is Kappa.

For illustration purposes, Kappa is calculated as follows for the matrix shown in Table B-1.

Step 1. 66 + 35 + 49 = 150

Step 2. (79 x 80) ÷ 200 = 31.60
(47 x 50) ÷ 200 = 11.75
(64 x 70) ÷ 200 = 22.40
65.75

Step 3. 150 - 65.75 = 84.25

Step 4. 200 - 65.75 = 134.25

Step 5. 84.25 ÷ 134.25 = .628 = Kappa

The obtained value of Kappa in this example (.628) is lower than we would like. If the two observers agreed on all of their ratings, Kappa would be 1.00. Thus, there is substantial room for improvement and additional training of the observers should probably be undertaken. An alternative might be to have the observers tally the frequencies with which the words yes, good, no, and bad are used, since this rating task requires no inferences on the part of the observer.

Ranking or rating. Kappa is an appropriate index of interrater reliability where entities, events, or behaviors are categorized. When numerical values are assigned to entities, events, or behaviors, however, some other type of index must be used. Consider the following example in which 14 classrooms were rated by two independent observers.

In this type of situation, the index of interrater reliability to use is the correlation between the two sets of ratings. You should note, however, that there are several types of correlation coefficients. You may be able to compute a "regular" (product moment) correlation coefficient using a pocket calculator or personal computer. That type of coefficient, however, makes assumptions about the properties of the rating scale that will probably not apply to your data. For this reason, we recommend converting the ratings to ranks and calculating a rank-orer correlation coefficient.

Table B-2

Ratings of Two Independent Observers

Rating

Classroom	Observer A	Observer B
1	6.0	6.5
2	7.3	6.8
3	2.2	3.8
4	4.5	5.0
5	9.7	9.0
6	4.3	8.7
7	8.0	7.7
8	7.9	7.9
9	7.1	6.4
10	7.3	6.4
11	6.6	6.4
12	8.5	7.7
13	8.1	9.2
14	6.4	7.0

Table B-3

Data Required for Calculating a Rank Order Correlation

	Observer A		Observer B			
Classroom	Rating	Rank	Rating	Rank	Rank Dif.	RD^2
1	6.0	11	6.5	9	2	4
2	7.3	6.5	6.8	8	-1.5	2.25
3	2.2	14	3.8	14	0	0
4	4.5	12	5.0	13	-1	1
5	9.7	1	9.0	2	-1	1
6	4.3	13	8.7	3	10	100
7	8.0	4	7.7	5.5	-1.5	2.25
8	7.9	5	7.9	4	1	1
9	7.1	8	6.4	11	-3	-9
10	7.3	6.5	6.4	11	-4.5	20.25
11	6.6	9	6.4	11	-2	4
12	8.5	2	7.7	5.5	-3.5	12.25
13	8.1	3	9.2	1	2	4
14	6.4	10	7.0	7	3	9

Table B-3 repeats the data from Table B-2 but includes the ranks of each observer's ratings, as well as the ratings themselves. It also includes a column of differences between the two observers' ranks and another column in which those differences are squared.

To calculate the kind of rank-order correlation coefficient we recommend, the following steps are required:

Step 1. Subtract the rank orders of one observer from those of the other observer.

Step 2. Square each of the rank differences obtained in Step 1.

Step 3. Sum the squared rank differences.

Step 4. Multiply the sum of the squared rank differences (from Step 3) by 6.

Step 5. Square the number of entities rated/ranked.

Step 6. Subtract 1 from the squared number of entities rated/ranked (from Step 5).

Step 7. Multiply the results of Step 6 by the number of entities rated/ranked.

Step 8. Divide the result of Step 4 by the result of Step 7.

Step 9. Subtract the result of Step 8 from 1. The "answer" is the rank order correlation coefficient.

Using the data from Table B-3, we would proceed as follows:

Step 1. (The rank differences are included in Table B-3).

Step 2. (The squared rank differences are included in Table B-3.

Step 3. The sum of the squared differences is 170.

Step 4. 6 x 170 = 1,020

Step 5. 14 x 14 = 196

Step 6. 196 - 1 = 195

Step 7. 195 x 14 = 2,730

Step 8. 1,020 ÷ 2,730 = .374

Step 9. 1 - .374 = .626 = rank-order correlation coefficient.

You should note that the presence of tied ranks (e.g. ranks 6 and 7 were tied for Observer A and both were given the rank of 6.5; ranks 10, 11, and 12 were tied for Observer B and all were given the rank of 11) causes the calculated rank-order correlation coefficient to be a slight overestimate of the true value.

There is a correction for tied ranks, but its effect is usually too small to worry about. For example, the correlation just calculated (.626) would drop only to .624 if the correction for tied ranks were applied.

Once again, the obtained interrater reliability coefficient is lower than would be desired. In this particular case, there is one sizeable difference between the rankings of the two observers (classroom 6 was ranked 3 by one observer and 13 by the other). One might wish to investigate this difference to see whether there was a clerical error or to find out why the observers had such different impressions. Normally, one would hope to obtain interrater reliabilities of .80 or higher.

The two procedures for calculating interrater reliability coefficients presented above were selected from a large sample of possible approaches because of their relative computational ease and freedom from strong assumptions. Some of the other approaches do have advantages, however, (e.g., the analysis of variance approach for assessing the interrater reliability of multiple raters). A more complete presentation of the topic is beyond the scope of the *User's Guide* but the interested reader is referred to any of the large number of statistics and evaluation design textbooks that deal with this issue.

APPENDIX C

FUNCTIONAL LEVEL TESTING

Why is Functional Level Testing Necessary?

Publishers of major achievement tests usually construct tests with several levels, each of which is designed to have suitable content and difficulty for children of specific ages or in specific grades. Publishers' recommendations about the level of test to use at a given grade level assume that both the curriculum and the achievement levels of the students to be tested are typical for that grade level. Their recommendations, however, will often be inappropriate for special populations, such as LEP students or students in need of compensatory education. Figure C-1 shows why it is important to test at a level that is appropriate for your students, when you are testing for the purpose of evaluating project effectiveness.

Suppose that you are considering two levels of a vocabulary test, A and B. These tests will sample a student's vocabulary and predict from each student's performance the size of his or her vocabulary. In your project, you have students T through Z. The publisher recommends test B for students of this age.

The lowest possible score on the level B test indicates a vocabulary of about 800 words. Students T, U, and V will all obtain this score on test B, despite the fact that their vocabularies range from about 25 words up to almost 2,000 words. The level B test is too hard for these students, and even if students T and U made good progress in vocabulary during the year, they could easily have the same score on a level B posttest that they had on the pretest.

The highest possible score on level B indicates a vocabulary of about 1,500 words. This is what student Z's score will show, although this student has a vocabulary of about 3,000 words. The test is too easy for this student. If he/she also took level B as a posttest, the score would not show any growth in vocabulary, because student Z obtained the highest possible score on the pretest.

If the level A test is used (one level below the recommended level) similar problems result. The test is still too difficult for student T, and it is too easy for students X, Y, and Z. None of these students will obtain scores that give a true picture of their vocabularies.

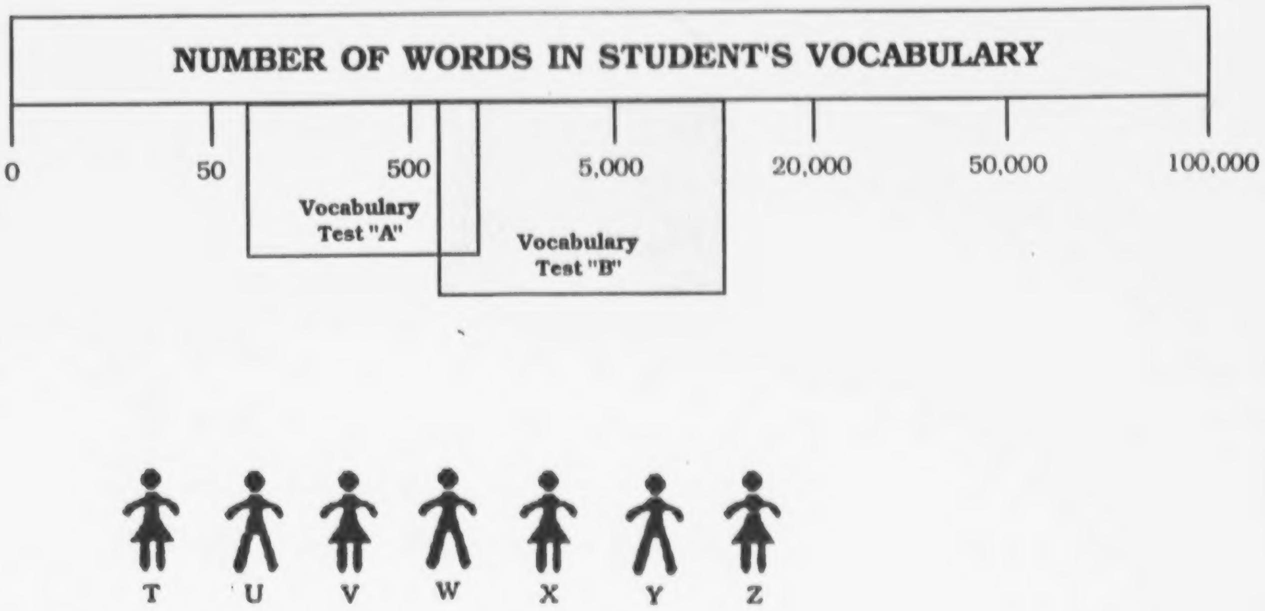

NUMBER OF WORDS IN STUDENT'S VOCABULARY
0
50
500
5,000
20,000
50,000
100,000
Vocabulary Test "A"
Vocabulary Test "B"
T
U
V
W
X
Y
Z

Figure C-1. Example of student scores on two vocabulary tests.

In our example, students U and V should be tested one level lower than the level recommended by the publisher, and student Z should be tested one level higher. Student T should be tested two levels lower. In some cases, you may need to move more than one level away from the publishers' recommendation in order to test a student at his or her functional level. It is more important to test students at an appropriate level so as to obtain an accurate gauge of their performance than it is to stay close to the publisher's recommended level.

How to Choose the Appropriate Test Level for Your Students

If you already administered a standardized test to your students, Chapter 5 contains rules of thumb you can use to determine whether the test you administered was at your student's functional level. You can also use scores from tests administered in the previous year to make similar judgments, keeping in mind that students' functional levels will increase over time. If you don't have access to scores from standardized tests for your students, there are other methods you may be able to use to determine whether a publisher's recommended test is at your students' functional level.

- *Locator tests.* The publisher of the test you want to use may publish a short locator test. You can administer this test to your students to identify the most appropriate level of the test to use.

- *Teacher judgments.* A teacher familiar with the test you are considering and with a student's classroom performance can often estimate the most appropriate test level.

- *Instructional materials.* The level of instructional materials students are using can be used as an indication of their functional levels.

- *Past classroom grades.* If a student's classroom grades are very low or very high, this might indicate that the test recommended for his/her grade level might not be at his/her functional level.

How to Use Scores From Functional Level Testing in the Gap Reduction Model

Using functional level tests with the gap-reduction model presents no problem. Simply use scale scores (sometimes called expanded standard scores) to perform all calculations. Scale scores are level-free. That is, if your student achieved a scale score of 60 on Level B of a test, you can assume that he/she would have achieved that same score on Level C. Simply remember to use the raw-score-to-scale-score conversion *table for the test level the students took* to look up their scale scores.

If you are using a "live" comparison group, the comparison group students may have been tested in level. In this case, you will convert their raw scores to scale scores using the in-level table. If you are using the test's norms as your comparison group, you will always use in-level scale scores. That is, if your *project group* is finishing third grade, you will obtain your *norm group* scale scores from the table for the test level recommended for students in the spring of third grade. Again, you will obtain your project group's scale scores from the table for the test level they took.

Summary

Functional level testing is sometimes necessary in order to obtain a good indication of students' abilities. If project students are tested at a level other than that recommended by the test publisher, the gap-reduction model can be implemented without any difficulty as long as the test publisher provides scale scores or expanded standard scores which can be used in gap-reduction calculations.

APPENDIX D

TRANSLATING TESTS

As stated in the text of this *Users' Guide*, translating tests is a difficult task that should not be taken lightly. There is a good chance that your efforts will be successful, however, if the test to be translated possesses the following characteristics:

- short questions.
- active voice.
- specific rather than general terms.
- no metaphors or colloquialisms.
- no vague words (probably, frequently, sometimes).
- no subjunctive mood.

Tests for use in kindergarten and the early elementary grades are likely to have most or all of these characteristics and thus represent prime candidates for translation. Tests designed for use in higher grades will provide more difficult challenges and, if alternatives other than undertaking translations are available, you would probably be well advised to adopt them.

According to the literature, the most common error made by "amateur" translators is that of being too literal. When word-for-word and construction-for-construction translations are made, the result is often stilted and ungrammatical in the "new" language. What is even worse, from a test perspective, is that word difficulty, and hence, item difficulty, may be changed dramatically when the most accurate word equivalencies are used.

It is, of course, essential that the translated material get the original ideas across. It is equally important, however, that it read as it would if composed by a native speaker and that the difficulty of the vocabulary used match that of the original. A translation that strives to meet these three goals will be far more successful than one that attempts to achieve exact literal correspondence.

"Back translation" is a technique that is often used to assess the success of a translation. When a person unfamiliar with the original text back translates, the back translation should bear close resemblance to the original. If "the spirit is willing, but the flesh is weak," after translation and back translation, comes out "the whiskey is good, but the meat is awful," something is amiss in the translation process.

The example just cited almost certainly arose from a too literal original translation. Such a translation would, no doubt, fail to convey the meaning of the original text. In fact, one might speculate that the original translation had a meaning more like the back translation than like the original text.

When there are significant discrepancies between the original text and a back translation, it may be helpful to paraphrase the original English until a more translatable (and back-translatable) version is found. One might, for example, paraphrase the original text to read, "people have good intentions, but lack the resolve to follow through on them." We suspect that a forward and back translation of this rephrasing would come closer to conveying the original meaning than the literal translation that back translated so badly. Of course, it is not always the original text that creates the problem. Sometimes it is the first translation. A second try may be much more successful.

A related problem may arise when the translator has a different ethnic background from the students who will be tested (or from the back translator). One would not wish to use a Cuban, for example, to translate a text to be used with students of Mexican heritage, as many words have quite different meanings for the two ethnic groups. Test questions could be phrased (accidentally) so that they would have one correct answer for one group but a different correct answer for the other group.

What has been said above can be summarized in the form of seven specific recommendations:

- Choose a translator who is fully bilingual and fluent in the particular dialect used by your target students.

- Impress upon your translator that you do not want a word-for-word translation. You want a translation that conveys the meaning of the original text, in a natural sounding and grammatically correct manner. Finally, you want the translation to have the same level of vocabulary difficulty as the original text.

- Choose a back translator who has the same qualifications as your "forward" translator.

- When discrepancies are found between the original text and the back translation, work with your original translator to identify the problem. Generate new translations directly from the original text as well as from paraphrasings of that text.

- Have the retranslations back translated.

- Continue the process until a translation is achieved that can be back translated successfully, i.e., so that the back translation conveys the same meaning as the original text.

- Review the translations for naturalness and correctness of grammar in the new language as well as for comparability of vocabulary difficulty. Make final adjustments.

All of these recommendations have been distilled from the relevant literature. They are, however, *highly* distilled. For more detailed guidance, the reader is referred to the article, *Constructing Matching Texts in Two Languages: The Application of Propositional Analysis* by Valdes, Barera, and Cardinas, that appeared in the Fall 1984 issue of the NABE Journal. It, and the references it cites, should be helpful.

APPENDIX E

MAJOR PUBLISHERS OF NATIONALLY STANDARDIZED ACHIEVEMENT TESTS

American Guidance Service, Inc.
Publishers' Building
Circle Pines, MN 55104
(612) 786-4343

CTB/McGraw Hill
Del Monte Research Park
2500 Garden Road
Monterey, CA 93940
(800) 538-9547
(800) 682-9222 (California)
(800) 649-8400 (Alaska, Hawaii,
and Foreign Countries)

The Riverside Publishing Company
8420 Bryn Mawr Avenue
Chicago, IL 60631
(800) 323-9540
(312) 693-0040 (Alaska, Hawaii and Illinois)

American Testronics, Inc.
P. O. Box 2270
Iowa City, IA 52244
(800) 553-0300

The Psychological Corporation
555 Academic Court
San Antonio, TX
(512) 299-1061

Science Research Associates
155 North Wacker Drive
Chicago, IL 60608
(312) 984-2195

APPENDIX F

TEST RELIABILITY

Test reliability is a complex topic that cannot be adequately covered here. Interested readers should consult one of the many psychometric textbooks that cover the topic. For our purposes, it is sufficient to note that highly reliable achievement tests yield scores (we will call them observed scores) that accurately reflect whatever real achievement differences exist among the individuals tested. Scores obtained from less reliable tests reflect these real differences less accurately.

Reliability is quantified in terms of *reliability coefficients* that have a theoretical range from 0 (totally unreliable) to 1 (perfectly reliable). Tests used in educational research and evaluation, however, usually have reliability coefficients that fall somewhere in the range from .60 to .95. These coefficients are related to the proportion of the differences among observed scores that result from real differences among the individuals tested. Observed-score differences that do not reflect real differences among the individuals tested result from measurement error.

The classic psychometric model specifies that observed variance (what we have been calling differences among observed scores) equals true variance (what we have been calling real differences among the individuals tested) plus error variance (what we have been calling measurement error). Since variance is quantified as squared standard deviations, we have:

$$s^2_{\text{observed}} = s^2_{\text{true}} + s^2_{\text{error}}$$

The amount of variance due to measurement error (error variance) is largely dependent on the characteristics of the test. It is thus relatively constant regardless of the particular group tested. On the other hand, the amount of variance due to real differences among the individuals tested (true variance) is clearly a function of how large those differences are.

If groups are homogeneous, the proportion of observed variance that is true variance will be relatively small. For heterogeneous groups, it will be relatively large. Since the error variance will be the same for the two types of groups, it follows that the reliability of a test will be lower if homogeneous groups are tested than if heterogeneous groups are tested.

The preceding point is extremely important to remember since the reliability of a test will be much lower when it is used with a bilingual project group than it will be when used with a nationally representative norming group. Since test publishers calculate their reliability coefficients from the scores of nationally representative samples, the coefficients they cite in their technical manuals do not provide a good indication of how reliable a particular test will be for your project group. Before you begin making reliability comparisons among candidate instruments, you should thus estimate what those reliabilities will be for your group.

To illustrate how this estimation process is done, consider the following example. We have a test with a national-sample reliability of .91 and a national-sample (observed score) standard deviation of 22.

Since we know that reliability equals $\frac{s^2_{true}}{s^2_{observed}}$

we can solve for s^2_{true} as follows:

$$.91 = \frac{s^2_{true}}{s^2_{observed}} = \frac{s^2_{true}}{(22)^2}$$

$$s^2_{true} = .91\ (484)$$

$$s^2_{true} = 440$$

Error variance then becomes 484 - 440 = 44.

Now, if we assume that our project group has only half the true variance of a nationally representative sample (a reasonable assumption), we have:

$$s^2_{obs} = 440/2 + 44$$

$$s^2_{obs} = 264$$

Then, the reliability of the test for our project group becomes:

$$\text{reliability} = \frac{220}{264} = .83$$

While the reliability reduction from .91 to .83 is not particularly dramatic, suppose that we are also considering a test that had a national-sample reliability of .75. The national sample would have the same true variance (440) as before, but the observed variance would now be 440 ÷ .75, or 586.7. Error variance would then be 586.7 - 440 or 146.7.

Now if we add the true score variance of our group (220) to the test's error variance we get an observed variance of 366.7. The reliability of the instrument for our group then becomes 220 ÷ 366.7 or .60.

As you can see, the reliability of the less reliable test drops substantially more (.15) than the reliability of the more reliable test (.80) when used with a homogeneous treatment group. The reliability difference between the two instruments is also larger for the homogeneous treatment group (.23) than it was for the national samples (.16).

At this point, we have not yet made clear how test reliability affects the accuracy of growth estimates. Estimating the reliability of the evaluation instruments for the specific group to be evaluated (as described above) is, however, a necessary first step. Once those reliability estimates have been derived, we can begin to assess their impact on the standard error of growth estimates.

The standard error of a growth estimate is given by the following formula:

$$\text{S.E.}_{\text{growth}} = \sqrt{\text{S.E.}^2_{\overline{X}_{\text{pre}}} + \text{S.E.}^2_{\overline{X}_{\text{post}}} - 2\, r_{\text{pre-post}}\, \text{S.E.}_{\overline{X}_{\text{pre}}}\, \text{S.E.}_{\overline{X}_{\text{post}}}}$$

where

$\text{S.E.}_{\overline{X}_{pre}}$ = the standard error of the pretest mean

$\text{S.E.}_{\overline{X}_{post}}$ = the standard deviation of the posttest mean

$r_{\text{pre-post}}$ = the correlation between pre- and posttest scores

Note that,

if $\text{S.E.}_{\overline{X}_{pre}}$ equals $\text{S.E.}_{\overline{X}_{post}}$ then:

$$\text{S.E. growth} = \sqrt{2\ \text{S.E.}^2_{\overline{X}} - 2\ r_{\text{pre-post}} \times \text{S.E.}^2_{\overline{X}}}$$

and

if $r = .5$,

$$\text{S.E. growth} = \sqrt{\text{S.E.}^2_{\overline{X}}} = \text{S.E.}_{\overline{X}}$$

As can be seen, if the two stated conditions are met, the standard error of a growth estimate is the same as the standard error of the mean pre- or posttest score.

For the sake of simplicity, we will proceed as if those two conditions are met. At the end of this discussion, we describe what happens if they are not met.

The standard error of a mean is given by:

$$\text{S.E.}_{\overline{X}} = \frac{s_{obs}}{\sqrt{N}}$$

As you may remember from the preceding discussion, s_{obs} increases as reliability decreases. Thus, the standard error of the mean (and, consequently, the standard error of the growth estimate) also increases as test reliability decreases.

Without going into the algebraic derivation, the percent increase in the standard error that results from choosing a less reliable test (L.R.T.) over a more reliable test (M.R.T.) is given by:

$$\% \text{ increase in S.E.} = 100 \times \frac{\sqrt{\text{rel.}_{MRT}} - \sqrt{\text{rel.}_{LRT}}}{\sqrt{\text{rel.}_{LRT}}}$$

Going back to our previous example, were we to select the instrument with a "local" reliability of .60 over the instrument with a "local" reliability of .83, we would increase the standard error of the mean by:

$$100 \times \frac{\sqrt{.83} - \sqrt{.60}}{\sqrt{.60}}$$

or 17.6%.

We have constructed Table F-1 to illustrate the percentage increase in the standard error of the mean that will be associated with selecting tests with specific lower reliabilities over tests with specific higher reliabilities. This table should be useful to you in estimating the impact on the accuracy of your growth estimates of test selections you may contemplate.

When using Table F-1, you should keep two things in mind:

- the percent increase in standard error is constant regardless of the size of your evaluation sample, but the absolute size of the increment may be inconsequential if your sample is large. Anything less than about a twentieth of a national-sample standard deviation is probably not worth worrying about.

- The standard error of a growth estimate may be either larger or smaller that the standard errors of the means that go into its computation. If the pretest-posttest correlation is .5 and the standard errors of the pre- and posttest means are unequal, the standard error of growth is less than .5--but it will be smaller if the correlation exceeds .5 (except as offset by differences between the standard errors of the pre- and posttest means).

Table F-1

Percent Increase in the Standard Error of the Mean When a Less Reliable Test is Selected

Rel. of Less Rel. Test	Reliability of the More Reliable Test															
	1.00	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.60	0.55	0.50	0.45	0.40	0.35	0.30	0.25
0.95	2.6%															
0.90	5.4%	2.7%														
0.85	8.5%	5.7%	2.9%													
0.80	11.8%	9.0%	6.1%	3.1%												
0.75	15.5%	12.5%	9.5%	6.5%	3.3%											
0.70	19.5%	16.5%	13.4%	10.2%	6.9%	3.5%										
0.65	24.0%	20.9%	17.7%	14.4%	10.9%	7.4%	3.8%									
0.60	29.1%	25.8%	22.5%	19.0%	15.5%	11.8%	8.0%	4.1%								
0.55	34.8%	31.4%	27.9%	24.3%	20.6%	16.8%	12.8%	8.7%	4.4%							
0.50	41.4%	37.8%	34.2%	30.4%	26.5%	22.5%	18.3%	14.0%	9.5%	4.9%						
0.45	49.1%	45.3%	41.4%	37.4%	33.3%	29.1%	24.7%	20.2%	15.5%	10.6%	5.4%					
0.40	58.1%	54.1%	50.0%	45.8%	41.4%	36.9%	32.3%	27.5%	22.5%	17.3%	11.8%	6.1%				
0.35	69.0%	64.8%	60.4%	55.8%	51.2%	46.4%	41.4%	36.3%	30.9%	25.4%	19.5%	13.4%	6.9%			
0.30	82.6%	78.0%	73.2%	68.3%	63.3%	58.1%	52.8%	47.2%	41.4%	35.4%	29.1%	22.5%	15.5%	8.0%		
0.25	100.0%	94.9%	89.7%	84.4%	78.9%	73.2%	67.3%	61.2%	54.9%	48.3%	41.4%	34.2%	26.5%	18.3%	9.5%	
0.20	123.6%	117.9%	112.1%	106.2%	100.0%	93.6%	87.1%	80.3%	73.2%	65.8%	58.1%	50.0%	41.4%	32.3%	22.5%	11.8%

APPENDIX G

QUASI-EXPERIMENTAL DESIGNS

This appendix deals with the estimation of project-related growth which is defined as the project group's total growth from pretest to posttest minus the growth it would have experienced had there been no project. In a true, randomized experiment, the total growth of the control group is taken as an unbiased estimate of the growth the project group would have experienced had there been no treatment. Then, the project group's total growth minus the control group's total growth yields our estimate of project-related growth (the project's impact).

Given this formula, the accurate estimation of treatment effects rests on the accurate measurement of project-group growth and the accurate measurement (or estimation) of control-group growth. Since, in bilingual education, randomly equivalent control groups are precluded by the legislative mandate that the neediest students be served, we have no choice but to abandon the task of estimating treatment effects or to use quasi-experimental estimation techniques in lieu of direct measurement of control-group growth.

In some bilingual education settings, quasi-experimental estimation techniques may be the logical choice. In other settings, however, it may simply be impossible to implement a quasi-experimental design in such a way that its results would be credible. In such instances, abandoning the estimation of treatment effects is the most prudent course of action.

Three quasi-experimental designs appear to have technical merit for evaluating bilingual projects. Each of them, however, has implementation requirements that many school districts will be unable to meet. These districts will thus have to confine their evaluation efforts to producing growth estimates and assessing gap reduction.

Overview of the Designs

The following paragraphs briefly summarize the three designs and their implementation requirements. If, after reading these summaries, you believe you may be able to implement one or more of the designs, you should go on to the more detailed descriptions presented later in this appendix.

The non-equivalent comparison group design. The first design to be discussed is the non-equivalent comparison group design. Despite the impression created by its name, the validity of this design hinges on the treatment and comparison groups being as nearly equivalent as possible. With substantially dissimilar groups, the assumptions underlying the design are not met and the statistical adjustments that can be made to correct for differences between groups are likely to be unsuccessful.

The primary difficulty associated with implementing this design is finding a nearly equivalent comparison group. A suitable group will almost certainly not exist within the same school as the treatment group. It may be possible to identify an acceptably similar group in another school in the same district. More likely, however, is the possible existence of a suitable comparison group in some other district where program entry criteria are significantly different from those in your district.

To illustrate, your district may have a small proportion of LEP students and may thus be able to serve Lau Category C students. Another district may have a much higher proportion of LEPs and may determine that it should devote all of its bilingual resources to Category A and B students. The Category C students in that district might constitute a reasonable comparison group--at least for the Category C students in your project. (Note: the comparability of the two groups might be even higher than it appears on the surface, since language proficiency tests tend to have quite low correlations with one another. This means that significant numbers of students identified as Lau Category C using one test would fall into Lau Categories B or D when tested with a different instrument.)

Of course the feasibility of implementing a non-equivalent comparison group design depends not only on the existence of a suitable comparison group but on your ability to obtain the necessary data (test scores, etc.) from its students. If you can do this, you should read the detailed implementation procedures presented below. If you cannot, you should consider one of the two remaining designs.

The grade-cohort design. What we refer to here as the grade-cohort design is really a hybrid that borrows some aspects of time series designs and some aspects of the non-equivalent comparison group design. Basically, it involves comparing the scores of students who have been in a bilingual project for some period of time with pre-entry scores obtained from other students who entered the project at the same grade level at which the project students were posttested. The difficult implementation requirement for this model is simply having enough students entering the program at different grade levels (and presumably without prior bilingual education experience) to provide a stable estimate of how the project students would be performing had they not been served by the project. An example may help to clarify this requirement.

Consider a school that receives large numbers of LEP students at all grade levels (perhaps because there is a large immigrant population). These students are all tested and then assigned to the bilingual education project. Suppose that, over the course of a few years, pretreatment scores were obtained for some 100 entering third graders.

Presumably there would also be a group of beginning third grade project participants--some of whom had been in the project one year, some two years, and some three years. By comparing the posttest scores of these children with the time-of-entry scores of the third-grade new arrivals, an estimate of project impact could be derived.

Although, in practice, the grade-cohort design should not be implemented in quite so simplistic a fashion, the example serves to illustrate its most significant features--including the need for a substantial amount of baseline data.

If you believe that this design could be implemented in your district, you should read the detailed implementation procedures for it that are presented below. If you feel that it cannot be implemented, you should consider the other two designs summarized here.

The regression-discontinuity design. Just as the validity of the non-equivalent comparison group design depends on a high degree of similarity between the project and comparison groups, the validity of the regression-discontinuity design hinges on the two groups having no overlap whatsoever on the pretest measure. All students scoring below some predetermined cutoff value on the pretest become members of the project group and all students scoring above it become members of the comparison group.

In addition to being able to establish and stick to a pretest cutoff score for assigning students to project and comparison groups, you must meet two additional requirements. The first of these factors is that the two groups must be drawn from a single population. The population might, for example, be students from homes where Spanish is the language spoken most of the time. Students from this population might then be assigned to project and comparison groups on the basis of their scores on an English language proficiency test. It would be inappropriate to implement the model with a mixed population of Hispanic and Anglo students if most of the students below the cutoff were Hispanics and most of those above the cutoff were Anglos.

The second requirement is that there must be a substantial number of students and a fairly wide range of pretest scores both above and below the cutoff. In theory, the model will work with small numbers and homogeneous sets of test scores. In practice, however, your estimate of project impact will be quite unstable under such circumstances.

If the conditions just described can be met in your district, you should read the detailed implementation procedures for the regression-discontinuity design presented below. If they cannot be met, you should consider the other two designs summarized here. If none of these can be implemented, you may skip the remainder of the material presented in this appendix.

Implementing the Non-Equivalent Comparison Group Design

One way to implement the non-equivalent comparison group design is to compare the two groups' posttest scores after they have been adjusted, by means of covariance analysis, for whatever differences existed between the groups at pretest time. Single or multiple covariates may be used but, in either case, the covariance analysis should be "corrected" for the unreliability of the pretreatment measures since it can be assumed that true rather than observed pretest scores mediate posttest performance.

The accuracy of the non-equivalent comparison group design hinges on the similarity of the two groups. If the two groups are identical at pretest time on all variables that correlate with posttest scores, then no statistical adjustment is required. If pretest scores are the only difference between the two groups, then using them as a single covariate provides an appropriate correction. But even if pretest scores are identical, the groups may differ on variables such as socioeconomic status, attitude toward school, academic aptitude and achievement motivation. To the extent that such variables are correlated with posttest scores, they must all be reflected in a "correct" statistical adjustment for initial differences between the groups. Any differences between the groups that remain unknown or unmeasured and that are therefore not included in the covariance analysis will result in a systematic misadjustment of posttest scores.

Given this situation, one might be tempted to measure everything he/she could think of to look for variables that (a) correlated with posttest scores, (b) discriminated between groups, and (c) had low intercorrelations with each other, as these are the three characteristics "good" covariates should have. Unfortunately, such "fishing" for significant relationships capitalizes on random error and thus results in a misadjustment for "real" between-group differences. The appropriate strategy is to consider the manner in which the two groups came into existence and the kinds of between-group differences that could be expected to result from the group-formation process. If the groups came from different communities, for example, one would suspect that there might be differences in socioeconomic status, in time spent in this country, in parent educational level, and even in the home language environment. Since all of these variables could be expected to correlate with posttest scores, all of them should be regarded as potential covariates. With project-size groups, however, probably no more than three covariates should be chosen. Additional covariates would be likely to add more measurement error than relevant true variance to the composite.

If there are pretest differences between groups, pretest scores should certainly be the first covariate chosen since these scores will correlate most highly with posttest scores. Choosing among the remaining alternatives might be as much a matter of cost and convenience as anything else. Whether students participate in free or reduced-price lunch programs, for example, might be a readily available piece of information, and, as such, a logical choice for a socioeconomic status covariate. If surveys of the home language environment have been conducted, the resulting information represents another "free" and potentially useful covariate. It might even be scaleable (in terms of, say, percent English spoken) as opposed to dichotomous (some English spoken vs. no English spoken).

Once the covariates have been chosen, the analysis can proceed. Data analysis with this design is, however, a nontrivial task and may require not only the use of a computer and a standard analysis program, but modifications to that program as well.

There are standard computer programs for analysis of covariance with multiple covariates in the BMD, SPSS, and SAS packages. None of them has a provision for introducing a reliability correction, however, and this calculation will have to be done either by hand or by modifying the existing program. The problem is further complicated by the fact that one or more of the covariates is likely to have an unknown reliability.

LISREL and EQS are two computer programs that appear able to deal with all these complexities. They are difficult and expensive to run, however, and would probably require your working with someone who had prior experience with them.

Although we are unable to offer a solution guaranteed to overcome the various difficulties inherent in the task of correctly adjusting posttest scores to compensate for initial differences between groups, it is probably useful to discuss the nature of the problem. In covariance analysis (leaving out several complicating details), a multiple correlation is calculated between the independent variables (covariates) and the dependent variable (posttest scores). This correlation is the primary determiner of the proportion of the initial difference between project and comparison groups that should be used to adjust the observed difference between mean posttest scores.

For the sake of simplicity, consider a case in which pretest scores were the single covariate and the comparison group outscored the project group by 10 points on the pretest but by only 5 points on the posttest. Apparently the treatment had the effect of improving the performance of the project group relative to the comparison group. But it is not appropriate to assume that the entire 10-point pretest difference would have carried over to the posttest had there been no project. Any part of the pretest difference that resulted from random error (unreliability) would not be expected to carry over. Another part of the difference which would not carry over to the posttest would be true variance that was unique to the pretest (not also measured by the posttest).

If there were no measurement error and the pre-and posttests measured exactly the same characteristic, it would be appropriate to use the full 10-point pretest difference to adjust the posttest difference, thus yielding a project-impact estimate of plus 5 points (-5 + 10 = +5). Under other circumstances, we would not use the entire 10-point pretest difference, and the estimate of project impact would be correspondingly less than 5 points.

The "regular" covariance analysis procedure would be to multiply the pretest difference by the "slope of the pooled, within-group regression line" and adjust the posttest difference by the result. Again making some simplifying assumptions, we can take this slope as equal to the pretest-posttest correlation--say .8. The regular covariance adjustment procedure would be to multiply the pretest difference of 10 points by this .8 and add the result to the difference between the two groups' mean posttest scores-- -5 +8 = +3. The covariance-adjusted estimate of project impact would thus be a gain of 3 points.

This regular covariance adjustment is appropriate when the project and comparison groups are formed by randomly assigning students drawn from a single population. The assumption, under such circumstances, is that measurement error is correlated with observed scores so that the higher of the two observed scores is relatively higher than it should be (compared to the true score for the group) and the lower score is relatively lower than it should be. The difference between the observed pretest scores of the two groups is thus somewhat larger than the difference between the corresponding true scores. Clearly, then, one should only adjust the posttest scores for that portion of the pretest difference that is due to the real differences between groups and not for that portion that is due to the measurement error (unreliability) inherent in the pretest scores.

When project and comparison groups are samples from different populations, rather than the same one, there is no reason to assume that any part of the observed difference between mean pretest scores is the result of systematic differences in measurement error. Measurement error will be correlated with observed scores within groups but not across them. Thus, the regular covariance adjustment will be systematically too low by an amount that is directly proportional to the amount of measurement error inherent in the pretest scores.

To remedy this problem, the pre-post correlation needs to be adjusted upwards to what it would be if the pretest scores were perfectly reliable. The formula for making this adjustment is quite simple.

$$\hat{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

where

$\hat{r}_{xy}$ = the estimated reliability-correlated pretest-posttest correlation.

r_{xy} = the observed pretest-posttest correlation.

r_{xx} = the reliability of the pretest for the groups tested.

If r_{xx} in the above equation were .79 and r_{xy} were .80, then $\hat{r}_{xy}$ would be .90. We would then adjust the posttest difference by nine-tenths of the pretest difference, or nine points:

$-5 + 9 = 4$

Our estimate of project impact is now +4 points.

It is hoped that the preceding discussion will give you some feeling as to how reliability-corrected covariance analysis "works" in adjusting posttest means to compensate for pretreatment differences between groups. Unfortunately, it does not provide a model for actually conducting such an analysis since it is unlikely that all of the simplifying assumptions we made would be met in any particular real-world situation. Without those simplifying assumptions, things get substantially more complicated.

It is possible to do an analysis of covariance--even a reliability-corrected analysis of covariance--by hand. The computations are very cumbersome, however, and you will almost certainly want to make use of a computer--especially if you use multiple covariates (which we recommend). You will probably need to seek assistance in carrying out the analysis regardless of whether you modify standard covariance analysis programs or employ the more complex causal modeling programs (LISREL or EQS).

After you have obtained your estimate of project impact we also recommend that you calculate 95% confidence limits around it. The standard error of that estimate should be included in or easily derivable from the computer printout. Multiplying that standard error by $\pm \underline{t}_{.05}$ will give you the upper and lower boundaries of your 95% confidence interval. Values of $\underline{t}_{.05}$ are provided in Table G-1.

Table G-1

Values of $t_{.05}$ as a Function of Sample Size (N)

N	$t_{.05}$	N	$t_{.05}$
2	12.706	19	2.101
3	4.303	20	2.093
4	3.182	21	2.086
5	2.776	22	2.080
6	2.571	23	2.074
7	2.447	24	2.069
8	2.365	25	2.064
9	2.306	26	2.060
10	2.262	27	2.056
11	2.228	28	2.052
12	2.201	29	2.048
13	2.179	30	2.045
14	2.160	40	2.023
15	2.145	50	2.009
16	2.131	60	2.001
17	2.120	100	1.984
18	2.110	∞	1.960

It was suggested earlier that it might be possible to find a more similar comparison group outside of the project group's school or even district than within those confines. Doing so, however, introduces a threat to the validity of your estimate of project impact that has not yet been discussed. The problem arises because the "regular" school program is likely to differ between schools within a district. Such differences are likely to be even larger between schools in different districts. While it is reasonable to assume that students in a bilingual project in a particular school would be in the regular program in the same school if there were no bilingual project, it is clearly less reasonable to assume that they would be in the regular program in some other school--particularly if that other school were in another district. Still, equivalence of regular programs is the assumption that is implicit in the design when it is implemented with a comparison group drawn from another school.

When the non-equivalent comparison group design is implemented with a comparison group drawn from a different school, it is essential that the credibility of that implicit assumption be investigated. If it can be verified that there is no difference between the program the comparison group students participated in and that which the project group students would have participated in had there been no bilingual project, the validity of the estimate of project impact is intact. If there are differences, they should be documented and their effects estimated. The credibility of the derived estimate of project impact hinges on a convincing argument that it could not be due to differences between the project and comparison schools' regular programs.

Implementing the Grade-Cohort Design

As mentioned earlier, the grade-cohort design involves comparing data obtained from students who have been in a bilingual project for some time with data obtained from new entrants at the same grade level. The baseline data may be obtained over a period of several years, but the assumption is always that the baseline data represent the level of performance that would be exhibited by program participants at the same grade level if they did not participate in the program. Although this assumption is basically the same as that employed by the two models described earlier in this appendix, it may, as we shall see, have quite a different meaning in the context of this design.

Once the data have been obtained and appropriately stratified (see below), implementing the model is extremely simple. Subtracting the mean pre-entry score of the comparison group from the mean posttest score of the treatment group yields an estimate of the project's impact. The standard error of the estimate can be readily determined (see any elementary statistics text), and confidence intervals can be constructed as described above for the nonequivalent comparison group design.

What would be ideal from a methodological point of view would be to have a district in which there was an influx of new families each spring (or fall) so that there were new children entering school *and* the bilingual project in time for spring (or fall) testing. Only children who had no prior bilingual education experience would be considered for the evaluation sample.

After a year in the project, the posttest scores of, for example, third graders who entered the project as second graders could be compared with the pre-entry scores of the children who entered as third graders. After two years in the project, the posttest scores of students who entered as first graders could be compared with the pre-entry scores of students who entered as third graders. Obviously the same general strategy would work for other grade levels and for longer "stays" in the project.

To avoid the possibility of the kind of self-selection bias that might result if students who were still enrolled in school after one, two, or three years were significantly different from students who stayed a shorter time, pre-entry data should be "stratified". This stratification would involve sorting pre-entry scores according to the length of time students remained in the school district. Thus the posttest scores of students who had been in the project for three years would be compared with the pre-entry scores of students who remained in the district three years. Second-year posttest scores would be compared with the pre-entry scores of students who remained in the district at least two years, and so on.

In the ideal scenario described above, all students would have entered school and the program at the same time and just before an assumed spring (or fall) testing. In a more realistic scenario, students would enter throughout the school year. Under these circumstances, pre-entry scores should have been collected at the time students entered the program.

If there were sufficient numbers of students, one could ignore (for evaluation purposes only) those who entered the program at times other than the month(s) when posttesting was done. Since large enough numbers are unlikely, however, the best strategy is to calculate a regression equation relating pre-entry scores to grade level (measured to the tenth of a month). The resulting equation should then be used to predict the scores that individual students would have obtained had they all been tested at the same time. The time for which the prediction is desired is the time at which students already in the program are posttested.

The regression (prediction) equation will have the form

$$\hat{Y} = Y + \frac{r_{gy} s_y G}{s_g} - \frac{r_{gy} s_y \overline{G}}{s_g}$$

where

$\hat{Y}$ = the predicted test score
$\overline{Y}$ = the mean of the observed test scores
G = the grade level for which the predicted test score is desired
$\overline{G}$ = the mean of the grade-level "scores"
r_{gy} = the correlation between the test scores and the grade-level scores
s_y = the standard deviation of the test scores
s_g = the standard deviation of the grade-level scores.

As noted earlier, it is important, when using the grade-cohort design, to compare the posttest scores of program participants with the pre-entry scores of *similar* children. The design, in fact, is a variant of what is typically called the posttest-only, control group design. As such, it rests on the assumption that the project and comparison groups are randomly equivalent--in other words, that they are both random samples from a single population.

The assumption of random equivalence may be questionable simply because the comparison group is created from historical data. In addition, the comparison group may include students from more than one year. Any population shifts that occurred between the pre-entry testing of the comparison group and the posttesting of the project group (such as the influx of new refugees) would clearly invalidate the assumption of random equivalence. An equally serious problem would arise from any systematic attrition from the project group that was unmatched in the comparison group.

Attrition will probably occur in the treatment group and will probably be non-random in nature--in other words, the students who drop out will differ systematically from the students who remain in the project. Students who remain in the project group should thus not be compared to all students who had pre-entry scores. It is *essential* that the pre-entry scores of students who dropped out of the program between their entry and the first posttest be eliminated from the comparison group before any of the calculations described above be undertaken.[1]

Once the proper (i.e., with the scores of dropouts eliminated) pre-entry mean score has been calculated, it should be subtracted from the mean posttest score of the treatment group. (Mean time in the program should also be calculated and reported for both groups to provide an indication of this comparability.) The difference between the two means is the estimate of project impact. Its statistical significance can be tested using the standard t-test for independent groups. In our particular case, the formula would take the following form:

$$t = \frac{\overline{Y_p} - \overline{Y}_c}{\sqrt{\frac{s_c^2}{N_c} + \frac{s_p^2}{N_p}}}$$

where

Y_p = the mean posttest score of the project group
Y_c = the mean of the pre-entry scores of the comparison group
s_c = the standard deviation of the comparison group's pre-entry scores
s_p = the standard deviation of the project group's posttest scores
N_c = the number of students in the comparison group
N_p = the number of students in the project group

---

[1] If we are looking at program effects of two, three, or more years of participation, it is essential to follow the same procedure of eliminating the pre-entry scores of students who dropped out of the program during the interval of interest.

The denominator of the above equation is the standard error of the project-impact estimate. The 95% confidence interval can be calculated (as described above) by multiplying the standard error by $\pm t_{.05}$.

As suggested earlier, estimates of project impact derived from the grade-cohort design *may not* be strictly comparable to estimates derived from the other two designs discussed in this chapter. In both of those two designs, expectations of how students would have performed without the project are derived from comparison groups that are observed over the course of a year during which they attended school but did not participate in a bilingual project. In the grade-cohort design, the no-project performance expectation is derived from pre-entry scores of students whose prior educational experience is likely to be unknown and may even be non-existent.

While the pre-entry scores of students who had been in school the previous year would be comparable to the posttest scores of comparison-group students at the same grade level in a non-equivalent comparison group design, that comparability would not hold for students entering school for the first time. A no-project expectation derived even in part from students who did not attend school the previous year would certainly be lower than one derived from school attendees. Estimates of project impact would be correspondingly larger.

Both types of no-project expectations are valid, but one tells us how project students would have performed if they had been in school but not in the project and the other tells us how project students would have performed had they neither attended school nor participated in the project. Evaluators using either the non-equivalent comparison group design or the regression-discontinuity design need not worry about this distinction. It may be important for evaluators using the grade-cohort design, however. Those evaluators should attempt to determine and report the proportion (if any) of the comparison-group students who did not attend school the year prior to their entry into the program. Then, if removal of these students from the comparison group will not make the group too small, they should be removed (but counted as members of the project group the following year). Better still, separate project-impact estimates should be calculated using the no-project-but-school expectation and the no-project-and-no-school expectation.

Implementing the Regression-Discontinuity Design

The regression-discontinuity design represents a special case of the non-equivalent comparison group design. Usually, the appropriate implementation of non-equivalent comparison group designs requires finding comparison groups that are as similar to the corresponding project groups as possible in all educationally relevant ways. In the regression-discontinuity design, the strategy is nearly the opposite. A group of students is subdivided into project and comparison subgroups so that there is no overlap whatsoever between their pretest scores. A cutoff score is established, and all students on one side of it are assigned to one subgroup while all students on the other side are assigned to the other subgroup.

One subgroup participates in the project while the other does not. Then both subgroups are posttested. Finally (in the simplest analysis strategy), within-subgroup regression lines are calculated. The distance between their intercepts with the cutoff score represents the impact of the project. Figure G-1 illustrates the regression-discontinuity design in a situation where regression is linear and where the project has had a substantial impact. At the level of the Figure G-1 example, the regression-discontinuity design is clear and easy to interpret. Unfortunately, real data rarely look like the illustrations one finds in statistics books. And when one fits linear regression lines to real data, the design may produce "pseudo effects." Experience with Chapter 1 evaluations was sufficient to demonstrate that a variant of the model, perhaps because of problems related to its implementation, tended to yield negatively biased estimates of project impact.

The design is particularly sensitive to test ceiling and floor effects, both of which produce curvilinear regression. Fitting linear regression lines to curvilinear data can produce either positive or negative pseudo effects, depending on where the cutoff score is located. To a large extent, however, the problem can be avoided by using curvilinear regression lines.

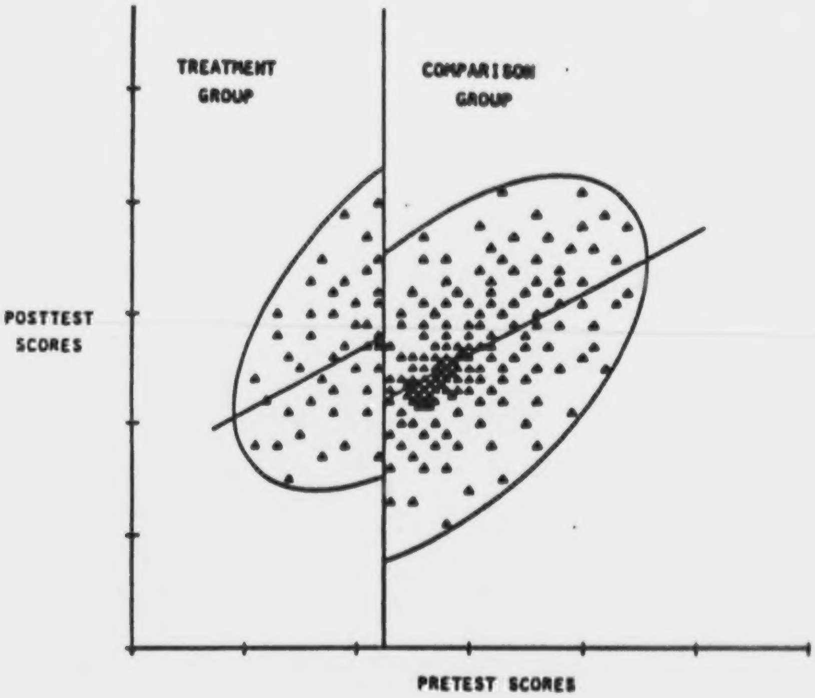


Figure G-1. The regression-discontinuity design (in its simplest form) with a substantial treatment effect.

Before discussing strategies for data analysis, it is perhaps useful to review the conditions under which the regression-discontinuity design will "work." The accuracy of a project-impact estimate derived from the regression-discontinuity design depends on how "tightly" the data points define the regression lines. That tightness, in turn, is a direct function of the sample size and the within-group correlations between pre-and posttest scores. Although, in theory, the design can be implemented with only a few students above and below the cutoff and a very low pre-post correlation, it would almost certainly be a waste of time and effort to do so.

recommendation is no more than a rule of thumb based on the probability that a "real" project effect will prove to be statistically significant, it is probably good advice. If pre-post correlations are high, fewer than 30 students might suffice, but if correlations are low, 30 students might not be enough. With respect to correlations, .40 has been recommended as a minimum. Although, again, this proposed cutoff value is purely arbitrary, trying to find a statistically significant project effect in the presence of a correlation much below .40 will probably prove fruitless--especially if sample sizes are small.

It is perhaps worth pointing out that correlations between sets of scores for any particular group of students depend to a large extent on the range of "talent" represented in the group. Two sets of scores will intercorrelate more highly in heterogeneous groups than in homogeneous groups. Ideally, then, one would like to have at least a moderate range of talent represented in both the above-cutoff and below-cutoff groups when implementing a regression-discontinuity design. The larger the range of talent, the higher the pretest-posttest correlation will be.

Another requirement, if the design is to work, is that a fixed cutoff score be established and that the assignment of students to groups be based strictly and entirely on that cutoff score. Breaking the rules is likely to introduce significant biases into effect-size estimates derived from the design. Analytic strategies have been developed for dealing with the problem of "fuzzy cutoffs" but we recommend that you use a fixed cutoff if at all possible.

Administrators and others may be reluctant to base participation on a fixed cutoff score on a single instrument. Fortunately, there is no need to do so. Multiple measures--including teacher judgments--can be used. The only requirement is that they be combined in such a way as to yield a single score. Once such a composite is developed, a cutoff score should be established for it.

At this point it is important to note that assignment to a bilingual project is certain to be based on some kind of assessment of English language proficiency. The regression-discontinuity design is thus clearly usable to assess the project's effectiveness in enhancing English language proficiency. Although it is less intuitively

The regression-discontinuity design is thus clearly usable to assess the project's effectiveness in enhancing English language proficiency. Although it is less intuitively sensible, the design can also be used to assess the effectiveness of the project's math, science, or social studies components (if the project has those components) even though the selection/pretest measure is not relevant to outcomes in those subject matter areas as long as the criterion of within-group correlations of .40 or larger is met.

The final requirement for valid implementation of the regression-discontinuity design is the most important and was alluded to earlier. The students above the cutoff score must be representatives of the same population as the students below it. If the majority of students below the cutoff belong to a particular ethnic minority group, the same proportion should hold for the group above the cutoff--otherwise there might be a discontinuity between the regression lines of the two groups that has nothing whatsoever to do with project impact.

As far as data analysis is concerned, the currently recommended procedure is to begin with a simple linear regression model such as that illustrated at the beginning of this discussion and then to reanalyze the data using curvilinear regressions of successively higher orders. After each analysis, the regression lines are plotted and visually inspected. After a number of analysis iterations (say 10) the size of the project-effect estimate, the amount of variance explained by the model, and the residual mean square values should all be plotted against the order of the regression equation used. The evaluator must then use judgment to select the equation that both provides a good fit to the data and makes intuitive sense. Unfortunately, there is no simple or mechanical way to select the best equation.

Although the analytic process sounds intimidating, and, indeed, manual computations are out of the question, explicit guidelines are available for using two packaged computer programs (SPSS or MINITAB). Only slight modifications of these guidelines are required for other statistical packages (SAS, BMDP, Datatest). The interested evaluator is referred to *Research Design for Program Evaluation: The Regression-Discontinuity Approach* by W.M.K. Trochim (Sage Publications, 1984) for additional information.

The computer printout should provide you with a standard error of the effect-size estimate. This statistic will enable you to compute a 95% confidence interval for your estimate following the same procedure described above for the nonequivalent comparison group design.

APPENDIX H

GAP-REDUCTION CALCULATIONS

The basic gap-reduction calculations are performed using pre- and posttest scores, and all students in the evaluation sample must have both. Before any gap-reduction calculations are done, however, several preliminary steps may be required.

Because the regulations require a 12-month testing interval, the posttest for one year will often also serve as the pretest for the next year. Each year's posttest may also serve a third function--that of determining (or helping to determine) which, if any, students will be exited from the project. If any students are exited at the end of the year and their posttest scores were considered in the exiting decision *these scores must be used in the subsequent year's evaluation even though the exited students did not participate in the project.* Those scores are essential for computing the correction that must be made for the regression effect bias.

If your project operated last year and served any students who were exited at the end of the year based at least partially on their posttest performance, you should calculate the regression-effect correction before you go any farther. The computational procedures are presented in Appendix J.

> Note: No correction is required if this year's participants were separately pretested--that is, if last year's posttest scores did not also serve as this year's pretest scores.

A second preliminary step that you may have to undertake involves estimating the scores that would have been obtained by students who were pretested late or posttested early had they been tested at the regular testing times. These estimated regular-testing-time scores are derived through a process of extrapolation. The computational procedures are described in Appendix I.

> Note: You only have to do the score extrapolations described in Appendix I if 10% or more of the students served by your project were either late enterers or early leavers (any students who participated in the project for less than 100 days need not be counted as served).

Once the preliminary steps have been completed, you can proceed with the gap-reduction calculations. To begin, you must compute the following statistics:

- The project group's mean pretest score.
- The project group's median pretest score.
- The comparison group's mean pretest score (if you are using normative data as your comparison group, the scale score corresponding to the 50th percentile).
- The standard deviation of the comparison group's pretest scores (if you are using normative data as your comparison group, the scale score corresponding to the 84th percentile minus the scale score corresponding to the 16th percentile divided by two).
- The project group's mean posttest score.
- The project group's median posttest score.
- The comparison group's mean posttest score.
- The standard deviation of the comparison group's posttest scores.

Once you have these statistics in hand, the next step is to decide whether to use the project group's mean or median pre- and/or posttest scores in future calculations. We recommend the following decision criterion: if the mean exceeds the median by more than a fifth of a comparison group standard deviation, use the median--otherwise use the mean.

Before proceeding--if the test you use has scale scores but some of your statistics are expressed in raw scores--you should convert the latter to the former.

Be sure to use the raw-to-scale-score conversion table that is appropriate for the test level you used.[1]

The Pretest Gap is the comparison group's mean pretest score minus the project group's mean/median pretest score divided by the comparison group's pretest standard deviation.

The Posttest Gap is the comparison group's mean posttest score minus the project group's mean/median posttest score divided by the comparison group's posttest standard deviation.

The Gap Reduction is the pretest gap minus the posttest gap.

The Comparison Group's (standardized) **Growth** is the comparison group's mean posttest score minus its mean pretest score divided by the square root of the average of its pre- and posttest squared standard deviations.[2]

The Project Group's (standardized) **Growth** is the comparison group's growth plus the gap reduction.

The Relative Growth Index is the project group's growth minus the comparison group's growth divided by the comparison group's growth and multiplied by 100 (to convert it to a percentage).

---

[1] To convert raw-score standard deviations to scale-score standard deviations, find the scale score corresponding to the raw-score mean plus one (raw score) standard deviation and the scale score corresponding to the raw-score mean minus one raw-score standard deviation. Subtract the latter from the former and divide by two.

[2] First square the comparison group's pretest and posttest standard deviations. Add them together and divide by two. Then, take the square root of the result.

If you wish to prepare a graphic display of your results (see Figure H-1), plot the comparison group's pretest score as zero. Next, plot the project group's pretest score minus zero as the pretest gap.

The next point to plot is the comparison group's pretest score, which is zero plus the comparison group's growth. Finally, plot the project group's posttest score as the comparison group's posttest score minus the posttest gap.

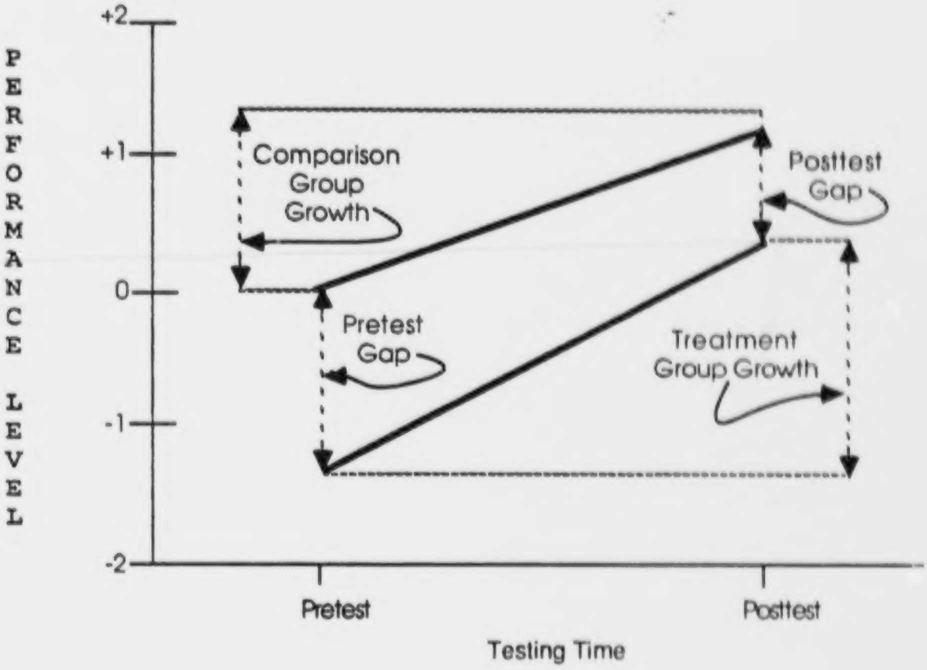

PERFORMANCE LEVEL
+2
+1
0
-1
-2
Comparison Group Growth
Posttest Gap
Pretest Gap
Treatment Group Growth
Pretest
Posttest
Testing Time

Figure H-1. Illustration of gap reduction.

APPENDIX I

EXTRAPOLATION PROCEDURES

If your project serves substantial numbers (more than 10%) of students who arrive after the pretest and/or leave before the posttest, you will need to obtain test scores from them in order to report representative achievement data. Under these circumstances, you will administer pre- and/or posttests to them at times that may be several months away from the dates you tested the rest of your project students. The test scores for these late-arriving/early-leaving students will show growth based on less than a full school year of project participation. In order to combine their scores with the scores of students who did receive a full school year of project services, you will need to extrapolate the scores of the late-arriving/early-leaving students. After extrapolation, their scores will represent an estimate of how the students would have performed if they had received a full year of project service and had been tested at the normal testing times. In order for such an estimate to be valid, the students must have received project services for a long enough period that one could expect the services to have made a difference. A good rule of thumb is to include scores only for students who have participated in the project for at least 100 school days.

Early-Leaving Students

The simplest case involves students who have "regular" pretest scores but who leave the project before the posttest would normally be administered. If forewarned of such students' departure dates, you can posttest them early. You then need to estimate what those students' scores would have been had they been posttested at the regular time.

If you are on a spring-to-spring testing schedule, the estimation procedure is a straightforward linear extrapolation (explained below). Fall-to-fall (and possibly others) testing schedules, however may involve projections that span the summer months. Since students are not in school and presumably not in the project during the summer months, it is not reasonable to assume that they would grow at the same monthly rate as they do during the school year. We follow common practice in this matter and make the assumption that the per-month cognitive growth rate during the summer is one-third what it is during the school year. The three summer months thus equal one school-year month and a full school year is made up of ten school-year months.

Consider the following example. A project participant is scheduled to move out of the school district at the end of February. To avoid losing his data, he is posttested on February 15th rather than the normal posttesting date of 30 April. Since the student was posttested 2.5 months early, his pretest-to-posttest growth reflects 7.5 *school-year* months of "treatment" rather than the 10 school-year months.

Our hypothetical student gained 9 test-score points (from 45 to 54) during his 7.5 months of project participation. This average monthly growth was thus 9 ÷ 7.5 or 1.2 points. Assuming he would have continued to grow at the same rate for the remaining 2.5 school-year months, he would have gained an additional 1.2 x 2.5 or 3 points. We can then simply add that 3 points to his early posttest score of 54. His estimated regular-testing-time score is thus 54 + 3 or 57.

If, in our example, a fall-to-fall testing schedule had been adopted and our student was posttested in mid-May instead of September 30th, we would have had to extrapolate his score over 4.5 *calendar* months. In terms of school-year months, however, we would have the same 2.5 month period as in the earlier example.

Late-Arriving Students

Late arriving students present a more difficult estimation problem. For one thing, it must be assumed that their pre-entry growth rate would have been less than their in-project growth rate. Thus we cannot simply make a linear projection backwards to estimate what their pretest score would have been if they had been pretested at the regular time.

For want of a better idea, we recommend assuming (as the computer program does) that a per-month growth during all months (summer *and* non-summer) prior to entry into the project was one-third of what was observed (per month) *during* project participation. To illustrate: Regular pre- and posttesting is accomplished at the end of April each year. Several new students who meet the project's entry requirements arrive in September, however, and are pretested on the 15th--4.5 calendar months after the regular pretesting date. Assume that their average pretest score was 37 and their average posttest score was 52. They thus grew 15 points during 7.5 months of project participation--2 points per month.

Had the student grown at one-third of their in-project rate during the 4.5 months prior to entering the project, they would have grown 2 x 1/3 x 4.5 or a total of 3 points. Subtracting those 3 points from the late pretest score of 37 yields an estimated regular-testing-time pretest score of 33.

Unfortunately, our score-projection task is not over yet. Even though we have a real, regular-testing-time posttest score, we need to estimate what that score would have been had the late-entering students been in the project for the full year (10 school-year months). To continue with our example, we must multiply the average growth rate of two points per month by the 10 months in a full school year. This gives us 2 x 10 or 20 points estimated total growth over the school year. Adding those 20 points to the estimated regular-testing-time pretest score of 33 gives us 53 as our estimate of what the late-entering students' mean posttest score would have been had they been in the project for the entire school year.

All later pretest and early posttest scores should be replaced by the corresponding regular-testing-time estimates before you begin your gap-reduction calculations.

APPENDIX J

CORRECTING FOR REGRESSION

Whenever the same test scores are used both to select students for participation in a project and as their pretest measures, the so-called regression effect bias will be introduced. You are likely to encounter this problem in two different situations.

The first situation arises when students are initially selected to participate (perhaps based on low scores on a language proficiency test). If those same scores are used as the students' pretest measures, they will be systematically distorted in a downward direction by the regression-effect bias. Measures of growth from pre- to posttest will, correspondingly, be spuriously inflated.

If, after the students are selected, they are administered a separate pretest (an alternate form of the selection test or preferably, a different test altogether), you need not worry about the regression effect bias. Administering a separate pretest is, in fact, the procedure we recommend *both* because it obviates any need to adjust scores for statistical regression *and* because language proficiency tests are not well suited for measuring growth over time.

The second situation where regression-effect biases will distort evaluation data is the one that is focused upon in Volume I of this *Users' Guide*. It arises when the posttest for one year of project participation also serves as the pretest for the subsequent year *and* some students are exited based, at least partially, on their posttest performance. In this situation, like the one described earlier, the posttest/pretest also serves as a selection test. Low scoring students are selected to continue in the project. Because of statistical regression their pretest scores will be lower than they should be and must be adjusted upwards to compensate for the regression-effect bias.

If no students are exited, or if the exiting decision is made without knowledge of the students' scores on the posttest, then there is no "selection on the pretest" and no significant amount of regression-effect bias. It would be nice if you could set your evaluation up so that you would not have to worry about the regression effect. Testing burden or other considerations may preclude that possibility, however. If so, then you should use the following procedures to calculate the appropriate correction.

Step 1. Assemble last year's posttest scores (this year's selection and pretest scores) for all students who were pretested (including exited students and dropouts).

Step 2. Calculate the mean and standard deviation of those scores.

Step 3. Exclude the exited students and repeat Step 2 for the "restricted group."

Step 4. Subtract the mean score of the restricted group (Step 3) from the mean score of the total group (Step 2).

Step 5. Divide the standard deviation of the total group (Step 2) by the standard deviation of the restricted group (Step 3).

Step 6. Identify that subset of students from Step 3 who were also posttested at the regular time.

Step 7. Calculate the pretest-posttest correlation for the students identified in Step 6.

Step 8. Enter Table J-1 with the ratio of the two group's standard deviations (from Step 5) and the pretest-posttest correlation calculated in Step 7. Read out the estimated correlation for the total, unrestricted group.

Step 9. Subtract the estimated total group correlation (from Step 8) from 1.0.

Step 10. Multiply the difference between the two group means (from Step 4) by the result of Step 9. The answer is the regression-effect correction.

At this point we recommend that you add the regression-effect correction to the pretest scores of each of the students in the restricted group (from Step 3). The full correction is only appropriate, however, when posttest scores were the sole criterion for exiting students. If posttest scores *and* teacher judgments (and possibly still other factors) entered into the exiting decision, the full correction will be excessive.

Under the circumstances just described, we recommend that you calculate separate gap reductions and RGI's using corrected and uncorrected scores and use the obtained values to "bracket" the correct values.

One final note: Exiting decisions are likely to be based entirely on English language proficiency. If this is indeed the case, you do not have to worry about the regression effect biasing gap-reduction or RGI computations for L1 proficiency or other academic subjects.

Table J-1
Table for Estimating the Total Group Pretest-Posttest Correlation

	r_{xy}-subgroup									
$\frac{SD_{x}\text{-group}}{SD_{x}\text{-subgroup}}$	.800	.790	.780	.770	.760	.750	.740	.730	.720	.710
1.50	.894	.888	.882	.875	.869	.862	.855	.848	.841	.834
1.48	.892	.886	.879	.873	.866	.859	.852	.845	.838	.831
1.46	.890	.883	.876	.870	.863	.856	.849	.842	.835	.827
1.44	.887	.880	.874	.867	.860	.853	.846	.838	.831	.824
1.42	.884	.877	.871	.864	.857	.849	.842	.835	.827	.820
1.40	.881	.875	.868	.861	.853	.846	.839	.831	.824	.816
1.38	.879	.872	.865	.857	.850	.843	.835	.828	.820	.812
1.36	.876	.869	.861	.854	.847	.839	.831	.824	.816	.808
1.34	.873	.865	.858	.851	.843	.835	.828	.820	.812	.804
1.32	.869	.862	.855	.847	.839	.831	.824	.816	.808	.799
1.30	.866	.859	.851	.843	.835	.828	.820	.811	.803	.795
1.28	.863	.855	.847	.839	.832	.823	.815	.807	.799	.790
1.26	.859	.851	.844	.836	.827	.819	.811	.803	.794	.786
1.24	.856	.848	.840	.831	.823	.815	.807	.798	.790	.781
1.22	.852	.844	.836	.827	.819	.810	.802	.793	.785	.776
1.20	.848	.840	.831	.823	.814	.806	.797	.788	.780	.771
1.18	.844	.835	.827	.818	.810	.801	.792	.783	.774	.766
1.16	.840	.831	.822	.814	.805	.796	.787	.778	.767	.760
1.14	.835	.827	.818	.809	.800	.791	.782	.773	.764	.754
1.12	.831	.822	.813	.804	.795	.786	.776	.767	.758	.749
1.10	.826	.817	.808	.799	.789	.780	.771	.762	.752	.743
1.08	.821	.812	.803	.793	.784	.775	.765	.756	.746	.737
1.06	.816	.807	.797	.788	.778	.769	.759	.750	.740	.730
1.04	.811	.801	.792	.782	.772	.763	.753	.743	.733	.724

	r_{xy} subgroup									
	.700	.690	.680	.670	.660	.650	.640	.630	.620	.610
$\frac{SD_{x\text{-group}}}{SD_{x\text{-subgroup}}}$										
1.50	.827	.819	.812	.804	.797	.789	.781	.773	.764	.756
1.48	.823	.816	.808	.801	.793	.785	.777	.768	.760	.752
1.46	.820	.812	.804	.797	.789	.781	.772	.764	.756	.747
1.44	.816	.808	.800	.793	.785	.776	.768	.760	.751	.743
1.42	.812	.804	.796	.788	.780	.772	.764	.755	.747	.738
1.40	.808	.800	.792	.784	.776	.768	.759	.751	.742	.733
1.38	.804	.796	.788	.780	.771	.763	.754	.746	.737	.728
1.36	.800	.792	.784	.775	.767	.758	.750	.741	.732	.723
1.34	.796	.787	.779	.771	.762	.754	.745	.736	.727	.718
1.32	.791	.783	.774	.766	.757	.749	.740	.731	.722	.713
1.30	.787	.778	.770	.761	.752	.744	.735	.726	.717	.707
1.28	.782	.773	.765	.756	.747	.738	.729	.720	.711	.702
1.26	.777	.769	.760	.751	.742	.733	.724	.715	.706	.696
1.24	.772	.763	.755	.746	.737	.728	.718	.709	.700	.690
1.22	.767	.758	.749	.740	.731	.722	.713	.703	.694	.685
1.20	.762	.753	.744	.735	.726	.716	.707	.698	.688	.679
1.18	.756	.747	.738	.729	.720	.710	.701	.691	.682	.672
1.16	.751	.742	.732	.723	.714	.704	.695	.685	.676	.666
1.14	.745	.736	.727	.717	.708	.698	.689	.679	.669	.660
1.12	.739	.730	.720	.711	.701	.692	.682	.672	.663	.653
1.10	.733	.724	.714	.705	.695	.685	.676	.666	.656	.646
1.08	.727	.717	.708	.698	.688	.679	.669	.659	.649	.639
1.06	.721	.711	.701	.691	.681	.672	.662	.652	.642	.632
1.04	.714	.704	.694	.684	.675	.665	.655	.645	.635	.625

	r_{xy} subgroup									
	.600	.590	.580	.570	.560	.550	.540	.530	.520	.510
$\frac{SD_{x\text{-group}}}{SD_{x\text{-subgroup}}}$										
1.50	.747	.739	.730	.721	.712	.703	.693	.684	.674	.665
1.48	.743	.734	.725	.716	.707	.698	.689	.679	.669	.660
1.46	.738	.730	.721	.712	.702	.693	.684	.674	.664	.654
1.44	.734	.725	.716	.707	.697	.688	.679	.669	.659	.649
1.42	.729	.720	.711	.702	.692	.683	.673	.664	.654	.644
1.40	.724	.715	.706	.697	.687	.678	.668	.659	.649	.639
1.38	.719	.710	.701	.692	.682	.673	.663	.653	.643	.633
1.36	.714	.705	.696	.686	.677	.667	.657	.648	.638	.628
1.34	.709	.700	.690	.681	.671	.662	.652	.642	.632	.622
1.32	.704	.694	.685	.675	.666	.656	.646	.636	.626	.616
1.30	.698	.689	.679	.670	.660	.650	.641	.631	.621	.610
1.28	.693	.683	.674	.664	.654	.645	.635	.625	.615	.605
1.26	.687	.677	.668	.658	.648	.639	.629	.619	.609	.598
1.24	.681	.671	.662	.652	.642	.633	.623	.613	.602	.592
1.22	.675	.665	.656	.646	.636	.626	.616	.606	.596	.586
1.20	.669	.659	.650	.640	.630	.620	.610	.600	.590	.580
1.18	.663	.653	.643	.633	.624	.614	.604	.594	.583	.573
1.16	.656	.647	.637	.627	.617	.607	.597	.587	.577	.567
1.14	.650	.640	.630	.620	.610	.600	.590	.580	.570	.560
1.12	.643	.633	.623	.614	.604	.594	.584	.573	.563	.553
1.10	.636	.627	.617	.607	.597	.587	.577	.567	.556	.546
1.08	.629	.620	.610	.600	.590	.580	.570	.559	.549	.539
1.06	.622	.612	.602	.592	.582	.572	.562	.552	.542	.532
1.04	.615	.605	.595	.585	.575	.565	.555	.545	.535	.525

	r_{xy} subgroup									
	.500	.490	.480	.470	.460	.450	.440	.430	.420	.410
$\frac{SD_{x\text{-group}}}{SD_{x\text{-subgroup}}}$										
1.50	.655	.645	.634	.624	.614	.603	.592	.581	.570	.559
1.48	.650	.640	.629	.619	.608	.598	.587	.576	.565	.554
1.46	.645	.634	.624	.614	.603	.593	.582	.571	.560	.549
1.44	.639	.629	.619	.608	.598	.587	.577	.566	.555	.543
1.42	.634	.624	.614	.603	.593	.582	.571	.560	.549	.538
1.40	.629	.618	.608	.598	.587	.576	.566	.555	.544	.533
1.38	.623	.613	.603	.592	.582	.571	.560	.549	.538	.527
1.36	.618	.607	.597	.587	.576	.565	.555	.544	.533	.522
1.34	.612	.602	.591	.581	.570	.560	.549	.538	.527	.516
1.32	.606	.596	.586	.575	.564	.554	.543	.532	.521	.510
1.30	.600	.590	.580	.569	.559	.548	.537	.526	.516	.505
1.28	.594	.584	.574	.563	.553	.542	.531	.521	.510	.499
1.26	.588	.578	.568	.557	.547	.536	.525	.515	.504	.493
1.24	.582	.572	.561	.551	.540	.530	.519	.509	.498	.487
1.22	.576	.566	.555	.545	.534	.524	.513	.502	.492	.481
1.20	.569	.559	.549	.538	.528	.517	.507	.496	.486	.475
1.18	.563	.553	.542	.532	.522	.511	.501	.490	.479	.469
1.16	.556	.546	.536	.526	.515	.505	.494	.484	.473	.462
1.14	.550	.540	.529	.519	.509	.498	.488	.477	.467	.456
1.12	.543	.533	.523	.512	.502	.491	.481	.471	.460	.450
1.10	.536	.526	.516	.505	.495	.485	.474	.464	.454	.443
1.08	.529	.519	.509	.499	.488	.478	.468	.457	.447	.437
1.06	.522	.512	.502	.492	.481	.471	.461	.451	.440	.430
1.04	.515	.505	.495	.484	.474	.464	.454	.444	.434	.424

	r_{xy} subgroup									
$\frac{SD_{x\text{-group}}}{SD_{x\text{-subgroup}}}$	.400	.390	.380	.370	.360	.350	.340	.330	.320	.310
1.50	.548	.536	.525	.513	.501	.489	.477	.464	.452	.439
1.48	.543	.531	.520	.508	.496	.484	.472	.460	.447	.435
1.46	.537	.526	.514	.503	.491	.479	.467	.455	.442	.430
1.44	.532	.521	.509	.497	.486	.474	.462	.450	.437	.425
1.42	.527	.515	.504	.492	.481	.469	.457	.445	.432	.420
1.40	.521	.510	.499	.487	.475	.464	.452	.440	.427	.415
1.38	.516	.505	.493	.482	.470	.458	.446	.435	.422	.410
1.36	.510	.499	.488	.476	.465	.453	.441	.429	.417	.405
1.34	.505	.494	.482	.471	.459	.448	.436	.424	.412	.400
1.32	.499	.488	.477	.465	.454	.442	.431	.419	.407	.395
1.30	.493	.482	.471	.460	.448	.437	.425	.414	.402	.390
1.28	.488	.477	.465	.454	.443	.431	.420	.408	.397	.385
1.26	.482	.471	.460	.449	.437	.426	.415	.403	.392	.380
1.24	.476	.465	.454	.443	.432	.420	.409	.398	.386	.375
1.22	.470	.459	.448	.437	.426	.415	.404	.392	.381	.370
1.20	.464	.453	.442	.431	.420	.409	.398	.387	.376	.364
1.18	.458	.447	.436	.425	.414	.403	.392	.381	.370	.359
1.16	.452	.441	.430	.419	.409	.398	.387	.376	.365	.354
1.14	.445	.435	.424	.413	.403	.392	.391	.370	.359	.348
1.12	.439	.429	.418	.407	.397	.386	.375	.365	.354	.343
1.10	.433	.422	.412	.401	.391	.380	.370	.359	.348	.338
1.08	.426	.416	.406	.395	.385	.374	.364	.353	.343	.332
1.06	.420	.410	.399	.389	.379	.368	.358	.347	.337	.327
1.04	.413	.403	.393	.383	.372	.362	.352	.342	.331	.321

	r_{xy} subgroup									
	.300	.290	.280	.270	.260	.250	.240	.230	.220	.210
$\frac{SD_{x\text{-group}}}{SD_{x\text{-subgroup}}}$										
1.50	.427	.414	.401	.388	.374	.361	.348	.334	.320	.307
1.48	.422	.409	.396	.383	.370	.357	.344	.330	.317	.303
1.46	.417	.405	.392	.379	.366	.353	.340	.326	.313	.299
1.44	.413	.400	.387	.374	.362	.348	.335	.322	.309	.295
1.42	.408	.395	.383	.370	.357	.344	.331	.318	.305	.292
1.40	.403	.391	.378	.365	.353	.340	.327	.314	.301	.288
1.38	.398	.386	.373	.361	.348	.336	.323	.310	.297	.284
1.36	.393	.381	.369	.356	.344	.331	.319	.306	.293	.280
1.34	.388	.376	.364	.352	.339	.327	.314	.302	.289	.277
1.32	.383	.371	.359	.347	.335	.323	.310	.298	.285	.273
1.30	.378	.367	.355	.342	.330	.318	.306	.294	.281	.269
1.28	.373	.362	.350	.338	.326	.314	.302	.290	.277	.265
1.26	.368	.357	.345	.333	.321	.309	.297	.285	.273	.261
1.24	.363	.352	.340	.328	.317	.305	.293	.281	.269	.257
1.22	.358	.347	.335	.324	.312	.300	.289	.277	.265	.253
1.20	.353	.342	.330	.319	.307	.296	.284	.273	.261	.250
1.18	.348	.337	.325	.314	.303	.291	.280	.269	.257	.246
1.16	.343	.332	.320	.309	.298	.287	.276	.264	.253	.242
1.14	.337	.327	.316	.304	.293	.282	.271	.260	.249	.238
1.12	.332	.321	.311	.300	.289	.278	.267	.256	.245	.234
1.10	.327	.316	.305	.295	.284	.273	.262	.252	.241	.230
1.08	.322	.311	.300	.290	.279	.269	.258	.247	.237	.226
1.06	.316	.306	.295	.285	.274	.264	.253	.243	.233	.222
1.04	.311	.301	.290	.280	.270	.259	.249	.239	.228	.218